# The effect of source reliability on the understanding of causal systems in primary and secondary school children

# The Effect of Source Reliability on the Understanding of Causal Systems in Primary and Secondary School Children.

A thesis submitted to
Birkbeck, University of London
for the degree of

PhD in Psychological Sciences

**GERMAINE SYMONS**

**22nd May 2019**

Department of Psychological Sciences

Birkbeck, University of London

## ABSTRACT

Individuals have excellent intuitive understanding of the physical world around them, evident from an early age. However, implicit understanding does not always transfer to explicit knowledge. The evaluation of source reliability is a crucial scientific reasoning skill that may assist in this transfer. Both adults and children have been shown to pay attention to source reliability, preferring higher reliability sources. However, assessments in children have generally used artificial manipulations of source reliability, and the degree to which younger children are showing epistemic awareness regarding potential source knowledge is unclear.

The study aims were to investigate the development of epistemic awareness in relation to what sources might know; to compare the developmental trajectory of implicit and explicit understanding of a familiar causal system; to enable a more direct comparison between the adult and child literature on source reliability; and to assess any role played by gender and language.

A more naturalistic task, a typical science class problem related to forces and motion, was employed. Six- to 17-year-olds were asked questions regarding their causal understanding, before and after receiving unexpected information from differentially reliable sources, and after carrying out an intervention, observing that the information was correct.

As predicted, participants who received information from high reliability sources were more likely to make correct predictions and explanations regarding the causal system. Participants who understood the causal system were more convinced than those who did not, and higher reliability source information increased conviction. Also, males made more correct predictions than females, although this could be confounded by age and SES differences. However, there were no age or language-related effects regarding source reliability, possibly due to demographic differences within the sample.

Future research looking at the role of source reliability in scientific reasoning should shift the paradigm into real-life environments, and include demographic and individual-differences measures.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

## ACKNOWLEDGMENTS

Firstly, I would like to thank the head teachers who assented to doing my studies in their schools. The staff in the schools were also very helpful both with recruitment and with providing a place to test. The children were also very keen to participate and made testing a pleasure.

I am grateful to my supervisor Mike Oaksford who was very supportive, and always knew what to say when the project seemed unmanageable. Thanks also to Andy Tolmie, my second supervisor, who never failed to answer an email, and was always an excellent source of information.

The staff and students in the Psychological Sciences Dept., where I work, were also excellent, providing ongoing support and encouragement especially in these final months where I was trying to both write up and do my job. In particular, Claire Santorelli, who was amazing. She provided ongoing, much needed support, and even did some excellent proof reading. Harish Patel is also worthy of special mention, as he built my apparatus, which was very sturdy and reliable, and performed with distinction throughout testing.

I would like to thank my children Harry and Loulou for their patience and understanding during this process; my father for his sterling effort on proof reading; my mother and sister Jessy for their support and help with formatting; and my other siblings, Tory, Jim and Sarah, for accepting my eccentricities during this time.

# 1 General Introduction

In the following sections, aspects relating to the development of scientific knowledge will be described. Firstly, causal understanding (section 1.1.1) will be discussed, where young children demonstrate implicit understanding of causal relations from a very young age. Secondly, the development of executive function (section 1.1.2) is briefly outlined as a factor that is likely to contribute towards scientific reasoning and understanding. Thirdly, scientific reasoning (section 1.1.3), and factors that are likely to influence scientific reasoning are discussed, including language (section 1.1.4), and science education (section 1.1.5). The impact of social economic status and gender are discussed throughout. Finally, the main focus of this thesis, research on source reliability understanding (section 1.2) is outlined, where source reliability plays a crucial role in learning about causal relations. The general introduction concludes with the aims of the thesis (section 1.3), an outline of the experimental paradigm (section 1.4), and the hypotheses (section 1.5).

## 1.1 Development of scientific knowledge

### 1.1.1 Causal understanding

#### 1.1.1.1 Definition

Moving from infancy to adulthood, an extraordinary understanding of the world is gained. The capability of representing the causal structure of the daily environment, and using that information to make accurate predictions regarding events in everyday lives emerges. In essence, an understanding of cause and effect is demonstrated, whereby even young children can predict and explain causal relations in many different contexts (Schulz & Gopnik, 2004).

The first indication of this understanding of causal relations was observed through the work of Baillargeon, Spelke, Carey, Gopnik and their colleagues, who produced a large body of research which suggested that young infants show appropriate understanding regarding physical objects and motion (see Spelke, Breinlinger, Macomber, & Jacobson, 1992, for review). The existence of this ability at such a young age suggests that they have an implicit understanding of the mechanisms that underlie these causal understandings, that are operational from very early on. It is likely this knowledge develops through interaction between innate core principles

and the environment (Spelke, 1994). Furthermore, these early intuitive understandings have led to the suggestion that people possess intuitive theories of different domains, such as physics, biology, and psychology (see Gerstenberg & Tenenbaum, 2017, for an extensive review of intuitive theories).

In the following, implicit understanding is thought to exist when children or adults make appropriate predictions regarding causal outcomes. They are not required to be able to accurately 'explain' why they made any particular prediction. This is in contrast to explicit understanding, where children or adults would be deemed to have explicit understanding when they provide accurate explanations regarding an appropriate prediction.

#### 1.1.1.2 CAUSAL UNDERSTANDING IN YOUNG CHILDREN

In an attempt to better understand the causal learning mechanisms that underpin causal understanding, and promote swift and effective understanding of causal relations in familiar environments, Gopnik and colleagues examined children's understanding of *unfamiliar* causal systems. An unfamiliar system allows systematic manipulation of the evidence children receive regarding causal structure, and examination of the causal inferences children draw as a result (Schulz & Gopnik, 2004). For example, the causal structure of these unfamiliar systems could be learned through observation, where the child observes contingency between events, and also through intervention, where the child intervenes on the system and observes different outcomes. Using simple causal systems which were easy to learn to operate allowed researchers to investigate causal understanding in children as young as two years of age.

For example, Gopnik, Sobel, Schulz, and Glymour (2001) developed a paradigm that made it possible to assess how children learned about novel causal relationships, without utilising possibly innate knowledge (such as addressing their understanding of the physical world), or involving explicit instruction (which is difficult for very young children to understand). They developed a 'Blicket detector' which lights up and plays music when a 'Blicket' (usually an object that is unfamiliar to the participant) is placed on it, but does not when other objects are placed on it. The Blicket detector allowed Gopnik et al. (2001) to assess what causal understanding young participants aged two, three and four years old could gain

from both observation and intervention. The majority of participants were from white, middle-class backgrounds. Participants showed causal understanding following observation of the contingency between objects (Blickets) that light the machine, and objects that do not. Gopnik et al. (2001) found that even very young participants would extend the term Blicket to newer objects that set off the machine, thereby suggesting that they believed that these newer objects played a causal role in activating the machine. That is, participants as young as two years of age appeared to make appropriate predictions regarding future causal events, following exposure to the actions of the Blicket detector. Furthermore, slightly older participants of three and four years old could use the information they had learned to intervene on the system and make the machine stop. Gopnik et al. (2001) concluded that participants appear to have generalised causal learning mechanisms that allow them to learn about new causal relationships, and modify causal systems as they learn new information.

Participants can learn new information about a causal system both through observation and intervention. When they use intervention, they change the causal system and observe the outcome. Schultz, Gopnik and Glymour (2007) found that four- to five-year-olds could use intervention on a novel causal system to determine causal relations, and use knowledge of causal structure to predict the outcome of an intervention. Furthermore, when left to intervene on the system on their own (or in a pair), participants were often able to learn the correct causal structure from evidence gathered during their own interventions. Schultz et al. (2007) also suggested that participants would be capable of learning real-world causal relations from their own interventions during play. Schulz, Goodman, Tenenbaum and Jenkins (2008), who collected data in a metropolitan museum, found that participants learned abstract rules regarding the cause and effect from sparse data, that was resistant to change in the face of anomalous evidence. If they attain these capacities, then in the case of playing, participants might well be able to form general causal rules regarding the environment in which they are interacting.

When looking at four- to five-year-olds' exploratory play, Cook, Goodman, and Schultz (2011) found that under certain ambiguous conditions, participants were more likely to perform an intervention to assess cause and effect, compared with more unambiguous conditions. Participants were also more likely attempt to assess

cause and effect when witnessing an outcome that was inconsistent with what they had previously observed, but not when it was consistent. When left to play freely with the causal system, they were more likely to engage in exploratory behaviour having witnessed an inconsistent outcome compared with a consistent one (Legare, 2012). Schultz and Gopnik (2004) investigated whether participants' ability to demonstrate causal knowledge in a variety of domains (such as biology, physics or psychology) was indicative of specific learning mechanisms, or a more general learning mechanism. Their sample included three- to five-year-olds, mostly from white middle class backgrounds. In the first three experiments, three- to four-year-old participants were presented with tasks in two domains, biology and psychology, and asked to identify what was the cause of the effect in structurally similar tasks in each domain. They found that both age groups made causal predictions using probabilistic information in a similar way across a range of tasks and domains. The results were consistent across domains. Furthermore, they found that four- to five-year-olds were capable of overriding their domain specific knowledge in light of contradictory evidence. Overall, Schultz and Gopnik (2004) did not find any differences in task response across domain. They suggested that (in this case) the context was not as important for causal inference, compared with probabilistic information they received. Participants are also capable of using information they learned in one domain to guide predictions in a second domain, when faced with a novel task. Schultz, Bonawitz and Griffith (2007) found that from three years old, participants were more likely to identify a cause for an effect that was consistent with their beliefs regarding a specific domain (here, biological or psychological) as within the same domain. However, four- to five-year-olds extended this by being more likely to identify a cross-domain cause for an effect, when it fit with the evidence, and less likely when it did not.

#### 1.1.1.3 Causal understanding in older children and adults

When considering developmental changes over time, in relation to causal learning, there is some evidence to suggest that younger participants are better at learning unexpected causal relations than older participants or adults. For example, Lucas, Bridgers, Griffiths, and Gopnik (2014), using the Blicket detector paradigm, found that four- to five-year-olds (from university affiliated preschools) paid more attention to the training data than did adults. The adults (undergraduate students)

appeared to be biased towards particular responses, where they persistently predicted only one Blicket was sufficient, even in the face of alternative evidence where two Blickets were required. This pattern of results was observable across different domains. Lucas et al. (2014) speculated that the adults' prior beliefs and experience influenced their understanding of how such causal systems might work, such that when they were exposed to unexpected changes in the causal systems, their biases made them reluctant to revise their beliefs. Whilst it is the case that Lucas et al. (2014) found this pattern of responding in a number of experiments, it is not clear that adults would exhibit similar behaviour in more real-life tasks. Adult participants were selected from undergraduate students who received course credit in an introductory psychology class for participating in these experiments. It is possible that a proportion of these students were devoting minimal cognitive resources to the task, and did not pay attention to the training period of the experiment. There were no individual differences analyses, nor were the adults debriefed on what they thought were the goals of the experiment, both of which may have shed some light on patterns of responding. In a conceptually similar study (data collected in a museum and a local preschool), when participants were presented with tasks which could have personal or situationally based explanations, four-year-olds appropriately inferred cause from the provided data. However, six-year-olds showed a bias towards personal based explanations, even when the data suggested a situationally based explanation (Seiver, Gopnik & Goodman, 2013). A similar argument to the above, that participants' prior beliefs and experience affected responding, was made to explain the age-related differences. This pattern has been found a number of times with younger participants, where the younger learners were better than the older ones. A potential explanation offered by Gopnik, Griffiths, and Lucas (2015) took a probabilistic model-based approach and argued that, if the learner has no prior beliefs regarding potential hypotheses, then they will not need much evidence to overturn one for the other. However, if the learner has stronger prior beliefs regarding one hypothesis, then the strength of evidence required to overturn that hypothesis would be much more substantial. If that was the case, one would expect age-related changes in performance when faced with unexpected evidence that contravened prior beliefs.

In light of this Gopnik et al. (2017) examined changes in causal learning from early childhood to adulthood. They included middle childhood and adolescence, as developmental periods that are not typically explored in this literature. Replicating the experimental procedure used in Lucas et al. (2014), they found a similar pattern of performance, whereby as the age of participants increased, their performance decreased. Four-year-olds were the best at learning unexpected physical causal relationships. Six-, and nine- to 11-year-olds did less well than the four-year-olds, but performed similarly. The 12- to 14-year-olds and adults also performed similarly, but performed less well than the younger age groups. Although this does not answer the question regarding the degree of attention paid to the task by adults, it adds support to the idea that greater experience impacts on prior beliefs, by biasing attention towards particular responses. If this is the case, one would expect an increase in bias with age, which was seen.

#### 1.1.1.4 Impact of socio-economic status and gender

There has been extensive research looking at causal learning and understanding in both younger children and adults. However, the majority of this research, particularly with children, has involved participants from English-speaking middle-class environments, with little attention paid to individual differences, such as gender, or socio-economic status (SES). This is the case for the research cited thus far as well, where gender and SES were rarely mentioned beyond description of the participant sample. The gender balance was usually reported as similar, and SES described as middle-class, and/or representative of the area within which the study was conducted. Ethnicity was usually described as representative of the area as well. In recent years, there has been a push towards making psychological science more representative of the human population (see Rad, Martingano, & Ginges, 2018). Inspired by this, Wente et al. (2017) observed that, to their knowledge, all of the studies looking at the development of causal learning used similar participant samples from higher SES environments in North America and Europe. To counter this, they conducted a study which included a relatively low income cross-cultural sample of Peruvian children and adults, as well as a low income SES sample of children from North America. They utilised the Blicket detector paradigm, following a similar procedure to Lucas et al. (2014). Both Peruvian four-year-old children (from private schools in low SES areas, designed to provide affordable education to

Peru's emerging middle class) and adults (Peruvian undergraduate college students) performed similarly to the higher SES North American equivalent; adults showed a bias towards particular responses, regardless of the evidence, which was not observed in the children. Children from lower SES environments (from a Head Start programme; Head Start promotes school readiness of children under five from low income families; U.S. Department of Health & Human Services, Office of Head Start, n.d.) largely responded similarly to children from higher SES environments, although the evidence was less clear for them. Wente et al. (2017) speculated that the lower SES children may have had some difficulties with inhibition and information processing that might have affected their responses. Furthermore, although these children come from low SES environments, their parents were interested enough in their education to participate in the Head Start programme, and it remains to be seen whether children with less parental support would perform equally well (Wente et al., 2017).

There appears to be little research looking at gender differences in the causal learning literature; it was not mentioned in the studies discussed earlier beyond noting that the gender balance of participants was similar. Galsworthy, Dione, Dale and Plomin (2000) found that gender accounted for only 3% of variance in verbal ability, and 1% in non-verbal cognitive ability in two-year-olds. When assessing the genetic and environmental contributions, they found that gender appears to influence early language development, but not non-verbal cognitive development. As such, one might not expect to find a gender difference in causal learning, given that gender differences tend to be quite small, particularly when looking at non-verbal cognitive ability. In one rare study that included gender and focused on causal reasoning, Bullock (1984) looked at children's understanding of causal mechanism when two objects moved in tandem. No gender difference in three- to five-year-old participants was found. However, even if gender differences in causal understanding and reasoning in younger children are minimal, that may not be the case for older children, especially once they have started school. It is possible that broader exposure to the environment around them, can influence prior beliefs, and that this may differentially affect genders in relation to their causal reasoning and understanding as suggested in Wente et al. (2017). This is because children participate in learning about science at school, where their explicit understanding

of causal relations is mediated by the classroom environment, where gender differences are frequently observed (and discussed in greater detail in section 1.1.3 on scientific reasoning).

#### 1.1.1.5 Adult causal understanding in complex environments

The evidence suggesting that children show causal understanding from a young age led Gopnik, Glymour, Sobel, Schultz, Kushnir, and Danks (2004) to suggest a theory of causal learning that states that children use specialised cognitive systems that allow them to create a learned, abstract, and coherent causal map of the world. Furthermore, they suggested that this map can be best understood in terms of directed graphical causal models. This is in keeping with the probabilistic models used to describe adult causal learning and understanding that have existed from the late nineties (e.g. Cheng, 1997; Pearl, 2000; Tenenbaum & Griffith, 2001). Initially, the majority of research looking at causal learning in adults focused on how people learn causal relationships with simple cause and effect examples (e.g. Cheng, 1997; Griffith & Tenenbaum, 2005; Shanks, 1995). However, the environments in which people operate involve multiple interacting variables, which dramatically increase the difficulty of inferring causal structure. This is because cause-and-effect relationships can only be inferred from observable cues such as contact, temporal order or covariation information (e.g. Kushnir & Gopnik, 2007; Lagnado & Sloman, 2006; see also Fernbach, Linson-Gentry & Sloman, 2007). One cannot *see* a cause and effect; one can only see that the action of one entity appears to 'affect' the action of another entity. In more complex environments it is necessary to understand how these individual cause-and-effect relationships interact with each other.

Seeking to address how adults learn the causal structure of more realistic, complex environments, Hagmayer and Waldmann (2000) asked German participants (no other participant information was given) to control a dynamic system by manipulating different variables to achieve a specific outcome. They were told nothing regarding how the system worked and were given a number of trials interacting with the system. They found that when the causal structure became more complicated, the adults showed little explicit understanding of the causal relationship between variables. However, they could make implicit, fairly accurate, predictions of singular events that took into account those very causal relationships. Furthermore, Steyvers, Tenenbaum, Wagenmakers, and Blum (2003) found that

there were individual differences in the causal inference strategies participants used. Undergraduate participants from a university in the USA, and web participants, were asked to learn to play a game about the communication networks of 'alien' mind readers, that could take different causal structures. In the first case, focusing on observation only, alien communications could follow one of two causal patterns, which participants were asked to identify after they had observed several examples of the alien communications. They found that the results for the undergraduate students and web participant were 'remarkably' similar. Based on observation alone, where participants only observed examples of alien communications, they found that participants were able to distinguish between competing hypotheses to varying degrees. Some participants learned like optimal Bayesians, integrating across trials; others functioned as one trial Bayesians, without integrating across trials; and some showed no causal learning at all. In the second case, participants were able to intervene to control one variable, the communication of a single alien, to examine its effect on the communication of other aliens, in order to determine the causal structure of the communication network. Participants became better at inferring causal structure, when they were able to intervene on the system, in comparison with observation alone. Furthermore, when they indicated they were unsure about causal structure, they subsequently chose interventions that would reduce their uncertainty. Individual differences in causal learning were also found by Osman and Shanks (2005). They found that not only did participants (undergraduate students from a UK university) differ in the way they weighted base rate information in a causal learning task, but that this also corresponded with performance in a decision-making task, where they treat base-rate information consistently across both tasks.

In summary, research evidence suggests that children are capable of learning causal relations, and making accurate predictions, from a young age, based on observation and intervention. It is possible that, as they develop, their learning is impacted by prior beliefs regarding the system at hand, where older children and adults perform less well when faced with unexpected evidence that conflicts with prior beliefs. However, for many, this understanding frequently does not appear to be explicit (even by adulthood). Causal understanding does not appear to be affected by gender, although there is little research explicitly taking gender into account. This is

also the case for SES, which may impact on causal learning and understanding, but the question is unresolved.

Other factors that may influence causal understanding are executive functioning, and language skills, which are discussed in the following sections (Executive Function – section 1.1.2; Language – section 1.1.4).

### 1.1.2 Executive functioning

#### 1.1.2.1 Definition

Executive functions (EFs) are a set of basic cognitive processes that allow the control and regulation of behaviour. The three core EFs commonly identified in the literature are inhibition control, updating, and shifting (see Miyake et al., 2000). Inhibition control refers to one's ability to inhibit a strong internal impulse, or external pull, to enable one to generate a required or appropriate response. This enables selective attention to specific stimuli, whilst disregarding or ignoring other stimuli (as both bottom-up and top-down processes; Diamond, 2013). Updating refers to the monitoring and updating of representations in working memory. This involves attention to and evaluation of incoming information as it relates to the relevant task, and then updating the information held in working memory according to that information (Miyake et al., 2000). This includes both verbal and visual spatial information (Diamond, 2013). Shifting refers to the ability to switch between alternative sets of mental operations, such as shifting between tasks, aspects of tasks, or perspectives (Miyake et al., 2000).

Higher level EFs such as reasoning, and problem-solving are likely to be underpinned by these primary lower level EFs (Diamond, 2013). For example, when attempting to identify specific causal relations, one might need to shift between potential hypotheses relating to potential causal relations, update one's belief in the face of new information, and inhibit prior beliefs in the face of new, contradictory, evidence.

The importance of executive functioning can be seen in its ability to predict school success, and other outcomes throughout life. In school, poor self-regulation negatively impacts on school success, and interventions designed to promote executive functions have been observed to counteract this (Blair & Diamond, 2008),

where executive functioning skills can be said to subserve successful self-regulation (Hoffmann, Schmeichel, & Baddeley, 2012). Moffitt et al. (2011) followed nearly 1000 children born in the same year in one New Zealand city for 32 years. They found that good self-regulation in childhood predicted many positive outcomes, including health, and personal finance.

#### 1.1.2.2 Factors related to the development of executive functioning

Children's ability to engage executive functions in goal-directed behaviour is predicted by age. For example, Freier, Cooper and Mareschal (2017) looked at cognitive control across three- to five-year-old participants from pre-schools local to the university in London, UK (information on SES was not included). They found age improvements, in that the older participants showed better cognitive control and were better able to produce an overarching goal. They did not find gender effects. Regarding older children, Lehto, Juujärvi, Kooistra and Pulkkinen (2003) looked at aspects of executive function in eight- to 13-year-old children from Finland (information on SES was not included) They found three inter-related factors similar to those of Miyake et al. (2000). Furthermore, they found age correlated with performance with most individual EF measures, as well as shifting and working memory (referred to as updating by Diamond, 2013). In a study that used a large representative national sample (from the USA) aged from five to 17 years, Best, Miller, and Naglieri (2011) found evidence to suggest that performance on the three core executive functions, inhibiting, updating, and shifting increases with age until at least 15 years. There was a big improvement around the ages of five to seven, and the increase slowed somewhat in adolescence.

#### 1.1.2.3 Impact of socio-economic status and gender

Another factor that possibly effects executive functioning (EF) is social economic status (SES). SES predicts EF, whereby children from lower SES backgrounds are likely to show poorer performance in tasks of inhibitory control, working memory, and other EF related tasks (e.g. Hackman, Gallop, Evans & Farah, 2015; Ng-Knight & Schoon, 2017). This SES difference can be seen when children start preschool (Welsh, Nix, Blair, Bierman, & Nelson, 2010) and was observed up to 10 years of age by Hackman et al. (2015). Negative impacts in adolescence have also been observed.

For example, lower SES has been associated with less behavioural inhibition (Spielberg et al., 2015).

The relevance of gender for executive functioning has also been investigated in younger children. Matthews, Ponitz, and Morrison (2009) used participants from a state wide longitudinal study investigating cognitive and social development during the transition from preschool to primary school in Wisconsin, USA. Just over 80% of participants were white, and 40% of participants' parents had Masters degrees. They found gender differences in five-year-olds self-regulation in the classroom, where teachers' ratings were more likely to favour girls over boys, although they found no gender differences in academic achievement in their study.

Yamamoto and Matsumura (2017) also looked at gender differences in five-year-old children, and combined direct measurement assessments of executive function tasks coupled with teacher ratings of behavioural self-regulation. They did not find gender differences in direct measurement assessments, but did in teacher ratings of self-regulation, with girls rated as having better behavioural self-regulation than boys. They speculated that teacher ratings could be gender biased, which may have also affected teacher ratings in Matthews et al. (2009).

Clark, Pritchard, and Woodward (2010) investigated gender in relation to EF as it relates to academic achievement, examining how well EF abilities predicted early mathematics achievement. They assessed children's developing executive function abilities at four years old, and looked at the relationship between these abilities and maths achievement at six years old. They found no relationship between gender and executive function. Gender was also not related to early maths ability.

Berthelsen, Hayes, White, and Williams (2017) found a similar pattern. They examined early self-regulatory behaviours at four to five years of age (as well as other child and family factors from early childhood), and how they were related to the development of executive function in adolescence (at 14 to 15 years), using a sample of approximately 5000 participants from Australia. They found higher attentional regulation at age four to five years for girls, as well as higher teacher ratings for approaches to learning at six to seven years of age. In contrast, they found that boys tended to score more highly on executive function tasks in adolescence. However, they speculated that this gender difference might be due in part to the

computer-based mode of assessment, where boys are likely to be more familiar with computer game playing. Differences in performance tend to be larger when computer-based, compared with when paper-based assessments are used (Jerrim, 2016). Ahmed, Tang, Waters and Davis-Kean (2018) also looked at the relationship between executive function and academic achievement from early childhood to adolescence in a US sample of around 1200 participants. Whilst not being a primary focus of their study, they reported that gender did not function as a significant predictor of EF, assessed at age 15, contrary to findings of Berthelsen et al. (2017). However, the tasks used to measure EF differed in the two studies, so it is not clear that the results are directly comparable. Although there is general agreement that EFs play a role in cognition, there is less consensus regarding which tests should be used in the assessment of specific EFs. If gender only has a small role in differences in cognitive development (Galsworthy et al., 2000; Ardila, Rosselli, Matute, & Inozemtseva, 2011), particularly when young, then one might expect task choice to affect the presence of gender differences, as speculated by Berthelsen et al. (2017).

In conclusion, executive function abilities are required for effective reasoning regarding causal relations. They improve with age, and are impacted by SES. Gender differences in executive function are less evident when the children are of primary school age but there is an indication that they may be more pronounced in adolescence.

### 1.1.3 Scientific reasoning

#### 1.1.3.1 Definition

Evidence of causal reasoning, particularly for very young children, usually stems from their ability to make correct predictions regarding causal relationships. However, the ability to explicitly reason about causal relations come somewhat later. While very young children might have an implicit understanding of causal relationships in a particular causal system, they may not have an explicit understanding of how a particular causal system works. Scientific reasoning is a more explicit activity, more closely related to language. It encompasses the myriad reasoning skills required in the generation, testing and revising of hypotheses and theory in an attempt to gain scientific understanding. Scientific reasoning skills are frequently used to gain a greater causal understanding of the world around us.

The literature looking at the development of causal reasoning/understanding and scientific reasoning tends not to overlap much, even though they largely address similar phenomena. The literature on causal understanding (outlined above) seeks to understand how we gain such sophisticated knowledge regarding the world around us, from such sparse data, and frequently uses Bayesian approaches to address this question (Tenenbaum, Kemp, Griffith, & Goodman, 2011). The field of scientific reasoning arose in part from the philosophy of science, which involved ongoing discussion regarding the "correct" scientific method. To this day, there is a large overlap between disciplines of philosophy of science and the psychology of (scientific) reasoning, where psychologists look to the philosophy of science to provide rational models with which to compare actual human reasoning. Furthermore, some of the more prominent models of normative scientific methodology take a Bayesian approach (Salmon, 1990; Talbott, 2008).

#### 1.1.3.2 Relationship between scientific and everyday reasoning

When considering how to define scientific reasoning, Klahr and Simon (1999) suggested that it is not qualitatively different from everyday reasoning, where scientific reasoning provides us with a systematic and cumulative construction of the world around us. Li and Klahr (2006) expanded on this concept and described scientific thinking by drawing on Einstein's characterisation of the relationship between scientific thinking and everyday thinking, where scientific thinking draws on the cognitive processes that underpin general problem solving. They suggested that scientific problem solving differs from everyday problem solving only in the more precise definitions of concepts and conclusions, along with more systematic choice of experimental material and greater logical economy.

Similarly, Kuhn (2010) also argued that scientific reasoning is akin to everyday reasoning, an activity most people engage in, rather than just professional scientists. Kuhn (2010) defined scientific reasoning as knowledge seeking, in that the person doing the scientific reasoning aims to enhance their knowledge, to gain scientific understanding of phenomena in the environment. The assumed parallel between scientific reasoning and everyday reasoning has led to much of the research on scientific reasoning looking at the ability of children and adults to reason normatively when faced with 'scientific' questions. Normative reasoning is

reasoning regarding the case at hand, according to the rules of scientific reasoning (see Stanovich, 2011, for an explication of those rules).

It is desirable to have a definition of scientific reasoning that is embedded in everyday reasoning as it is more likely to reflect how scientists actually reason. One of the problems that bedevilled earlier models of scientific method, which sought to describe models of how science should be done, such as the falsificationist method, was that historical examples of successful scientific progress frequently do not map on to the outlined method (Salmon et al., 1992). For example, the failure of Newtonian mechanics to predict the orbit of Uranus was not considered disconfirmatory evidence, as earlier outlines of falsificationist method suggested it should have been. Instead, an auxiliary hypothesis was suggested to explain the orbit (Anderson & Hepburn, 2016; Salmon et al., 1992). Newtonian mechanics were not considered to be disconfirmed until a better theory came along, the theory of relativity, which could explain the phenomena under question, as well as the perihelion of Mercury, which Newtonian mechanics could not (Anderson & Hepburn, 2016). The failure of the falsificationist method to explain the history of science has led to it being rejected by contemporary philosophers of science (Oaksford & Chater, 2007), where the conclusion is that normative models of the scientific method should be able to explain past scientific progress.

This problem has also imbued the psychology of reasoning, which looked to philosophy to provide optimal, rational models of reasoning with which to compare everyday reasoning (Oaksford & Chater, 2007). Although there was a belief that human reasoning was inherently rational, people have persistently provided illogical responses in some tasks, such as the Wason selection task (see Oaksford & Chater, 2007 for a review). In the Wason selection task, participants are presented with four cards, with a number on one side and a letter on the other. Participants are then asked which card(s) they would turn over to establish whether a given rule was true or false. According to the falsificationist method, participants should seek to disconfirm the rule, but very few participants applied the nominal 'normative' strategy, suggesting that people's hypothesis search strategies were inherently illogical. Ultimately, new models of the scientific method came to the fore, most notably taking a probabilistic approach, which made much more sense of people's strategies in the Wason selection task, as well as other tasks where people's

responses failed to conform to the normative response (see Oaksford & Chater, 2007, and Stanovich, 2011) for extensive discussions of this issue).

Along with the argument that any definition of scientific reasoning *should* map on to everyday reasoning, there is also evidence to suggest that people *do* treat every day and scientific arguments in a similar way. Corner and Hahn (2009) asked undergraduate participants to evaluate the strength of science and non-science arguments. They found that the evaluation of both appeared to be determined by the same factors, such as diagnosticity of evidence – strength of evidence and reliability of the source of evidence – where stronger evidence and more reliable sources led to higher ratings regarding the strength of the argument for both science and non-science arguments. The main difference between the two was that ratings for science arguments were sometimes more polarised. Science arguments were rated as very strong with high reliability sources and strong evidence, and very weak for the converse, low reliability sources and weak evidence. The difference was not as extreme for non-science arguments. Corner and Hahn (2009) suggested that this may be due to the way in which science is currently taught as a collection of objectively drawn facts, based on evidence. As such, a scientific argument from an unreliable source, and lacking evidence, might seem particularly unconvincing, compared with a non-science argument which does not suffer from the burden of such expectations.

In conclusion, an influential strand of current thought in the philosophy of science holds that models of the scientific method should be able to describe events within the history of science - actual scientific practice – which supports the idea that a definition of scientific reasoning should draw from everyday reasoning (Kuhn, 2010; Li & Klahr, 2006). If a definition of scientific reasoning does not track how people actually reason, then it is likely not to explain many aspects of the scientific process (Oaksford & Chater, 2007). Definitions such as this seek to place scientific thinking within the grasp of the majority, as opposed to the elevated few. It also suggests that scientific reasoning is in principle teachable, in that it is an extension of reasoning that is used every day.

#### 1.1.3.3 FACTORS RELATED TO THE DEVELOPMENT OF SCIENTIFIC REASONING

Regarding the development of scientific reasoning, there has been a parallel drawn between children's reasoning, where they seek to gain understanding of the world around them, and scientific reasoning. As presented in section 1.1.1 on causal reasoning, children appear to have some knowledge regarding our physical environment from a very young age (e.g. Carey & Spelke, 1996; Gopnik, Meltzoff, & Bryant, 1997; Gopnik et al., 2001; Spelke, 1994). The existence of these early intuitive understandings has led to the suggestion that people possess intuitive theories of different domains, such as physics, biology, and psychology (see Gerstenberg & Tenenbaum 2017 for an extensive review of intuitive theories) and some have claimed that these intuitive theories are similar to scientific theories. That is, both contain a system of interrelated concepts, governed by causal laws that dictate the relationships between the concepts. They suggested that, if this is the case, then cognitive development is like theory change in science (Carey, 1985; Gopnik et al., 1997; Schulz, 2012), although others have argued the contrary (Fuller, 2011). However, Kuhn (2010) proposed that, although children construct implicit theories regarding their environment which are updated in light of new evidence (e.g. Gopnik et al., 2001), this process cannot be described as scientific reasoning as the process occurs outside conscious awareness. As scientific reasoning develops, using conscious thoughts and language, the process of theory revision becomes more explicit, whereby theory and evidence are considered in relation to one another. To be a competent scientific reasoner, one would need to have the set of skills required to allow the generation, testing and revision of theories in light of evidence. One should also be able to reflect on the process of acquiring and revising knowledge (Kuhn, 2005).

Several reviews have summarised the comparison between children's reasoning and scientific reasoning (Carey 1985; Gopnik et al., 1997; Schulz, 2012). Klahr and Simon (1999) pointed out that the evidence is problematic as children exhibit flaws in their (scientific) reasoning, such as failing to design controlled experiments, failing to revise belief in the face of new evidence, failing to discriminate between theory and evidence, and so on (see Silar & Klahr, 2012, for a review of general misconceptions in scientific reasoning). For example, Klahr, Fay and Dunbar (1993) compared nine-year-olds, 11-year-olds, undergraduates (science or engineering

majors), and community college students (non-science majors; the latter two groups comprise participants with and without scientific training). The child participants came primarily from academic and professional families, i.e. middle-class environments. Participants were taught how to use a programmable robot. They were then given a new operation, given a potential hypothesis, and asked to discover what the actual rules were. Potential hypotheses were always incorrect but either plausible or implausible. They found that adult participants were much better than children at choosing experiments that decreased the number of potential hypotheses. Adult and child participants also differed in response to the plausibility of the initial hypothesis. When the hypothesis was plausible, both adults and children sought to elucidate features of the hypothesis. However, when the hypothesis was implausible, adult participants tended to propose alternative hypotheses, and conduct experiments that discriminated between them. They were much better than the children at discovering implausible rules. Children tended to focus on plausible rules, and struggled to detect implausible but correct hypotheses from the data. Nine-year-olds showed poorer performance than 11-year-olds. Overall, adults were better than children at evaluating hypotheses and designing experiments.

The development of cognitive skills is related to the development of scientific reasoning. Although young children can use perceptual evidence to guide their understanding of causal relations (see section 1.1.1 on causal understanding), it is thought that children need to understand that belief and knowledge can be based on perceptual evidence. It is claimed that this develops along with the development of theory of mind, in that one needs be able to recognise a protagonist can hold false beliefs in relation to the external world (Kuhn, Cheney and Weinstock, 2000; Morris, Croker, Masnick, & Zimmerman, 2012). By age five, children are likely to show basic scientific reasoning prowess such as utilising evidence to determine whether or not a hypothesis is knowable. Fay and Klahr (1996) asked five-year-old participants from middle-class homes in the USA to choose which set of objects could produce a particular outcome, where the problem was determinable sometimes, and indeterminable at other times. They found that most of the children could solve the determinable problems, and some of them could solve the indeterminable problems (in that they identified the problem as indeterminable). There were notable

individual differences in responses, and only a small proportion of participants made optimal responses in the more difficult conditions.

Koerber, Sodian, Thoermer and Nett (2005) examined four-, five-, and six-year-old German middle-class participants' ability to assess their understanding of the hypothesis evidence relationship, by examining what they understood about the impact of new evidence on beliefs. The children were asked about a character's beliefs following evidence that contradicted the character's prior beliefs, or faked evidence. All age groups showed high accuracy when required to revise the belief of the character, although around 25% of four-year-olds were excluded due to failing the control questions designed to assess their ability to complete the task appropriately. The authors suggested that four-year-olds' performance may be limited by memory and attention problems. These problems may also explain why four-year-olds did not do so well in the faked evidence task. They would have to hold concurrent representations of the evidence, where memory and attention might impact. Furthermore, in a second experiment, Koerber et al. (2005) also showed that five-year-old participants' ability to evaluate evidence was affected by the prior beliefs, although even here their ability to assess the hypothesis evidence relationship was still above chance. This finding is in keeping with similar findings in the causal understanding literature, where prior beliefs impact on understanding (see section 1.1.1 on causal understanding). This pattern of response is also seen with adults (Chin & Brewer, 1998), who gave different types of responses to data that conflicted with non-scientist adults' prior beliefs. Koerber et al. (2005) suggested that their, and other similar findings (e.g. Sodian, Zaitchik, & Carey, 1991; Ruffman, Perner, Olson, and Doherty, 1993), indicated that even participants as young as four years old demonstrate basic scientific reasoning skills.

Mayer, Sodian, Koerber, and Schwippert (2013) looked at the impact of EF on scientific reasoning (specifically at inhibition control and problem solving) using paper and pencil tests with 10-years-olds from Germany, with a sample balanced by gender. No relationship was found between scientific reasoning and inhibition. They did find a relationship with problem solving (plus IQ, reading comprehension, and spatial abilities). It is possible the lack of relationship with inhibition is due to the nature of the test. Mayer et al. (2013) used a variant of the Stroop colour-words task to measure inhibitory control. Participants did not need to set aside prior beliefs

about causal relations in this test, so it is not clear that the task used represents the kind of inhibitory control one might need when doing a scientific reasoning task. When they controlled for all cognitive abilities included in the study, scientific reasoning performance was best predicted by problem solving and reading comprehension.

However, children frequently confuse theory and evidence even in simple cases. For example, when four- and six-year-olds, and adults, were asked to explain why a certain event occurred, participants frequently responded with their theory regarding why the event occurred, rather than reporting the evidence (Kuhn & Pearsall, 2000). Participants (no more information was provided) were shown a sequence of pictures indicating an event with a particular outcome, such as two runners competing in a race. Participants were asked to indicate the outcome and justify the knowledge. Four-year-olds were more likely to provide a theory as to why the outcome came about, rather than report the evidence that pertained to the outcome. In this example, when indicating who they thought had won, they would refer to the fast shoes, worn by the winner (theory regarding the outcome), as opposed to the trophy the winner was holding (evidence of the outcome). Six-year-olds showed a much better understanding and adults always provided an evidence based response in these relatively simple tasks. There was a small improvement at each age group when they were given a prompt towards an evidence based response (adults were at ceiling; reported in Kuhn and Pearsall, 2000). Notably, the difference between the findings of Fay and Klahr (1995), and Koerber et al. (2005) and those of Kuhn and Pearsall (2000) is that the latter required explanatory responses. That is, the participants were required to explicitly understand and be able to explain the hypothesis evidence relationship that led to their response. Koerber et al. (2005), on the other hand, asked participants to select between different potential responses, a far simpler task linguistically, that did not require extensive explicit understanding to participate. It could be argued that these findings mirror the findings of causal understanding in young participants - that participants show good basic understanding of hypothesis evidence relationships, but at younger ages, understanding is not fully explicit. Taking the definition of scientific reasoning outlined above - the reasoning skills required in the generation, testing and revising of hypotheses and theory in an attempt to understand the world around us - then

these children (Fay & Klahr, 1996; Koerber et al., 2005) showed the underpinnings of scientific reasoning. Until such time as the process occurs inside conscious awareness, and children can provide explicit explanations for their responses, they cannot be considered to be competent scientific reasoners.

#### 1.1.3.4 CONTROL OF VARIABLE STRATEGIES

One crucial skill in scientific reasoning is the ability to design experiments that adequately assess the question at hand. In particular, to design unconfounded experiments, and draw appropriate inferences from the outcomes. This is frequently referred to as a control of variables strategy (Chen & Klahr, 1999). The control of variables strategy is a method of creating experiments in which a single contrast is made between experimental conditions (this also known as a 'fair test'). It also includes the ability to make appropriate inferences based on the outcomes of these unconfounded experiments.

Chen and Klahr (1999) looked at how children from seven to 10 years old, from two private elementary schools in the US, acquired and understood the procedural and logical aspects of using the control of variables strategy to do experiments. Participants interacted with three tasks in the physical domain (springs, inclines, sinking) that had multiple variables that could affect the outcome, with a specific question that needed to be addressed. For example, the incline angle, height of starting gate on the incline, surface on the incline, and type of ball could vary, and participants were asked what factors determined how far the ball will roll down a ramp. Some participants were given control of variables strategy training. Probe questions were also asked: firstly why they chose the test they did; and secondly whether or not they could tell that the variable they were testing made a difference, and why. Participants who had (explicit) training plus probe questions (implicit training), were compared with participants who had no training plus probe questions, and a control group who had no training and no probe questions. Following training on one task, they were assessed on that task, and then were introduced to two more tasks, to see whether any transfer effects could be observed. It was found that a small percentage (15%) already knew the strategy and used it from the start. Furthermore, with appropriate explicit training, participants were capable of learning control of variables strategy, with transfer within the same domain, similar domains, and 10-year-olds could transfer the strategy to other more

remote domains. Participants in the probe only group (implicit training), and the control group showed much less use of the control of variables strategy and no difference between them. However, although explicit training boosted performance, participants did continue to use ineffective strategies in some cases. Chen and Klahr (1999) concluded that, although children can learn aspects of the appropriate procedural and logical underpinnings of the control of variables strategy from a young age, grasp of these strategies improves with age. Furthermore, they suggested that direct instruction is an effective way to educate children with regards to control of variables strategy.

Kuhn and colleagues used a microgenetic approach, whereby the participant engaged in similar tasks over multiple sessions such that changes in strategy could be observed (see Kuhn, 2010; Kuhn and Dean, 2005). For example, Kuhn and Dean (2005) asked 11- to 12-year-olds to participate in a control of variables task. They sought to consider efficient methods other than direct instruction to develop children's inquiry skills through engagement and practice. However, this cohort was different from Chen and Klahr's (1999) study as they were academically low performing children from an urban public school serving a lower income population. This cohort was chosen as they often showed limited progress in developing inquiry skills over time, compared with students from more advantageous backgrounds. The study was designed to prevent students from engaging in ineffective testing (such as manipulating multiple variables at once), which frequently occurred if students failed to formulate an appropriate question. Kuhn and Dean (2005) speculated that children were attempting to discover the effects of all variables at once, and intervened by suggesting to participants that they should find out about one thing at a time. All participants who received the suggestion indicated a focus on only one variable. Participants in the control condition focused on only one variable around 11% of the time, and 83% of the time they intended to find out about three or more variables in a single comparison. Furthermore, they found participants in the experimental group made many more appropriate causal inferences compared with the control group, in both the original and a new context, even with a cohort that frequently underperforms compared with peers from more advantageous backgrounds. Dean and Kuhn (2007) also found little difference in long-term performance of 10- to 11-year-olds from diverse SES

backgrounds in the USA, in a direct comparison between direct instruction and practice, and practice only. However, they found that children's use of the control of variables strategy improved and became more embedded if direct instruction was accompanied by sustained practice involving problems requiring the strategy.

Kuhn (2007a) argued that control of variable strategies do not encompass the whole of scientific reasoning. For example, not only does one want to identify the effects of a single variable among many, but also to reason appropriately about simultaneous effects of multiple variables (once the individual effects have been identified). Kuhn (2007a) sought to identify whether children's poor performance at multiple variable prediction was related to poor control of variables strategies, or could be attributable to uncertainty regarding the structure of the causal system at hand, and vacillation over the nature of different possible effects. Participants were from an independent school affiliated with the University, and approximately nine- to 10-year-olds. Kuhn (2007a) found that participants failed to utilise the information they gained regarding the effects of different variables, to make accurate predictions regarding multiple variable effects. A similar finding was also found with adults. Kuhn (2007b) investigated the everyday reasoning of the average adult (as opposed to undergraduate students) on familiar topics. It was also found that people with college degrees performed better than people without (Kuhn, 2007b), which may suggest that extensive training in skills related to scientific reasoning are required to perform well in more complex tasks, such as making multi-variable predictions. As with explanations for poor performance above, Kuhn (2007b) also argued that prior beliefs regarding the variables at hand may have impacted on performance.

Further research looking at strategies used to identify underlying causal mechanisms suggested that the use of effective strategies improves with age, with adults much more likely to use an effective strategy (Chen & Klahr, 1999; Klahr et al., 1993). This is not to say that younger participants do not use effective strategies, just that they are more likely to use less adequate ones. However, the majority of participants of all ages are likely to improve with practice (Kuhn & Dean, 2005), suggesting that developing good scientific reasoning skills involves an introduction to the necessary skills *plus* practice of those skills.

Research with older children and adults suggests that the continued development of scientific reasoning becomes even more complex as the complexity of the questions at hand increase. In more complex cases, even adults confuse theory and evidence, frequently relying on explanation to support claims (Brem & Rips, 2000; Kuhn, 2001). This phenomenon is dependent on context, and disappears among abler university students (see Kuhn, 2001, for summary). Adults are also likely to have problems integrating theory and evidence, and there is a great range of scientific reasoning ability in both children and adults, although adults are almost always better than children (Klahr et al., 1993; Kuhn & Pease, 2006).

#### 1.1.3.5 EPISTEMOLOGICAL UNDERSTANDING

When thinking about how things are known more generally - epistemological understanding – Kuhn et al. (2000) sought to illustrate individual differences in epistemological understanding, as well as changes over time. They suggested that mature epistemological understanding is likely to reveal understanding of the coordination of the subjective and objective dimensions of knowing. They identified four categories of knowing: *realist* (assertions are copies of an external reality; knowledge is from an external source, and is certain); *absolutist* (assertions are facts that are correct or incorrect in their representation of reality; knowledge is from an external source, and is certain, but not directly accessible, possibly resulting in false beliefs); *multiplist* (assertions are opinions freely chosen by and accountable only to their owners; knowledge is generated by human minds, and is uncertain); and *evaluativist* (assertions are judgements that can be evaluated and compared according to criteria of argument and evidence; knowledge is generated by human minds, and is uncertain but susceptible to evaluation). Mature epistemological understanding would be demonstrated by understanding at the evaluativist level. People who engage in fully specified scientific reasoning regarding phenomena in the world could be said to be reasoning at the evaluative level, whereby scientists evaluate and compare assertions (hypotheses, theories) according to criteria of argument and evidence.

Kuhn et al. (2000) compared 10-, 13-, and 17-year-old mostly middle-class, white students to undergraduate students from a high-ranking university, as well as three adult groups, one group chosen to be of comparable intellectual ability to the undergraduate group (working full-time, and doing a business MBA; predominantly

white), and the other of seemingly lesser ability (community college students, of primarily Hispanic ethnicity). The final group was an expert group of PhD candidates. Their goal was to consider the influence of age, intellectual ability and life experience on epistemological understanding. They developed a measure of epistemological understanding, asking a series of questions which enabled them to assess what epistemological understanding category the participants reached, in a number of different domains (judgements of personal taste, ascetic judgements, value judgements, judgements regarding truths about the social world, judgements regarding truths about the physical world). Participants were given a pair of contrasting statements from two interlocutors. They then had to indicate whether they thought one or both statements could be 'right'. That was followed up with a second question they needed to evaluate, that if both statements could be 'right', could one statement be better or more right than the other.

Kuhn et al. (2000) suggested that very young children begin as realists, becoming an absolutist during early science education years. Of particular relevance here were judgements regarding truths about physical world. They found around a third of 10-year-olds showed a predominantly absolutist level. This decreased slightly, but did not disappear in all other groups except for the expert group. A similar but opposite pattern was observed, whereby around a fifth of 10-year-olds showed a predominantly evaluativist level, which increased with age, to around 40-45% for the undergraduate students and adult groups, aside from the expert group, all of whom reached this level.

One issue with developmentally based claims is that they do not fully concur with the children's early demonstration of basic scientific reasoning skills, which one could argue show the beginnings of an evaluativist level of epistemological understanding (Fay & Klahr, 1996; Koerber et al., 2005; see also causal understanding section 1.1.1.2). For example, Koerber et al. (2005) found that four-to six-year-olds were capable of revising beliefs in light of new evidence, when provided with a task that was linguistically simple, only requiring implicit understanding. Kuhn et al. (2000) used a more complex assessment, which required relatively sophisticated linguistic understanding. It is possible that some differences in performance across age could reflect developmental changes in language ability and understanding (see section 1.1.4 on language), as opposed to a transition

through the described levels of epistemological understanding, particularly as they relate to the evaluativist level.

#### 1.1.3.6 Explanation

One skill that is necessary for developing good scientific reasoning is the ability to generate appropriate arguments following predictions. Kuhn (2005) suggests that the central goals of science education should be to teach children both enquiry skills, and argument skills. Enquiry skills emphasise the processes of enquiry - asking questions, the generation and interpretation of data, and drawing conclusions. In primary schools in the UK the focus of science education is largely on the development of enquiry skills (5-11 years; Department for Education, 2015). In secondary school (11-16 years; Department for Education, 2015), there is an expanded focus on enquiry skills, and more emphasis on argumentation skills. These include being able to provide and modify explanations; taking into account the relationship between the data, predictions and hypotheses; and being able to draw appropriate conclusions based on evaluation of evidence and argument. Providing explanations is crucial to the practice of science, where argumentation takes a central role. The findings of science, when presented for the evaluation of others, are largely provided in the form of written explanatory arguments, explaining the relationship between theory and evidence.

Explanation also plays a central role in our everyday reasoning. They are generated to help us make sense of the world, and they assist us in the exchange of beliefs, and making decisions. Lombrozo (2006) claimed that explanation also assists us in our causal understanding, whereby explanation constrains causal inference by reducing the range of potential mechanisms to those consistent with prior beliefs regarding the causal mechanisms. Certainly, children engage in explanation when utilising causal reasoning to make sense of real life events, and do so frequently, even at age two to three years. Hickling and Wellman (2001) examined the content of explanations that four two-and-a-half to five-year-olds (75% white, 75% middle class) gave or asked for in everyday conversation. They found that the explanation is focused on the entity targeted, and the explanatory mode of causal reasoning, plus the relations between these elements. They concluded that these children had appropriately constrained, yet flexible causal reasoning. Legare and Lombrozo (2014) also found that explanation appeared to promote causal learning and

generalisation. Their participants came from preschools in a metropolitan area in south-west USA. They were gender balanced and primarily Euro-American and middle-class. Participants were shown how to use an unfamiliar causal system, and then were either asked to observe the machine, or provide an explanation regarding how it works. The experimenter then removed a part of the causal system, and the child was then asked which part would make it work again, choosing from a selection of parts. Legare and Lombrozo (2014) found three- to five-year-olds showed better understanding of the causal system when they had been asked to provide an explanation. Furthermore, they also found age-related improvements, where the four- and five-year-olds were better than the three-year-olds. Similarly, Walker, Lombrozo, Williams, Rafferty and Gopnik (2017) found that, when making novel generalisations, five-year-old children who explained (as opposed to reported) observations using apparatus similar to the Blicket machine were more likely to favour hypotheses with broader scope. The participants were recruited in both university preschools, and a local museum. Demographic information was not collected. However, both preschool students and the museum visitors were approximately 60% white. Participants who generated explanations were also more likely to prefer hypotheses that concurred with their prior beliefs, as has been found previously (e.g. Gopnik et al., 2017). In order to investigate developmental differences in the relationship between explanation and exploration, Legare (2012) involved two- to six-year-old participants (predominantly Euro-American and middle-class.) Explanations regarding consistent and inconsistent outcomes were compared. For inconsistent outcomes only, the type of causal explanation differentially predicted exploratory behaviour. If children provided causal function explanations, they engaged in more hypothesis testing type behaviours. Legare (2012) did not find any age-related differences, suggesting that the task was too simple to identify developmental differences. However, the sample size in each age group was quite small so one particpant's performance could have more of an impact. This, coupled with the fact that individual differences were not taken into account, may have clouded any age-related differences that might exist.

However, providing explanations can also be problematic. In research with older children, Kuhn and Katz (2009) investigated whether self-generated explanations were always beneficial. Participants were nine to 10-year-olds, from a school with

around 50% white students, and a broad range of SES backgrounds. The study was part of a science class, where, over multiple sessions, participants investigated a causal system with the goal of identifying causal relations. Some participants provided explanations regarding their beliefs of the causal mechanisms, on a regular basis, whereas others did not. They found that participants who did not provide explanations did better on a transfer task compared with participants who did; they were more likely to ignore relevant evidence following explanation of causal relations. Kuhn and Katz (2009) suggest that explanation is reinforcing prior beliefs, which have already been observed negatively impacting on scientific reasoning tasks (e.g. Gopnik et al., 2017; Koerber et al., 2005). Explaining has also been observed leading to over generalisations in young adults. Williams, Lombrozo, and Rehder (2013) found that undergraduate participants who explained were more likely to seek broad patterns, which hindered learning when patterns involved exceptions, in category learning tasks. Similarly, Berthold, Röder, Knörzer, Kessler, and Renkl (2010) found that German tax-law students who provided conceptually based explanations were more likely to show good conceptual knowledge, compared with participants who did not. They provided more detail, and elaboration in their explanations. However, providing explanations negatively impacted on the acquisition of procedural knowledge. It is possible that providing repeated explanations functioned as a form of retrieval practice, which is well-known strategy for learning (Brown, Roediger, & McDaniel, 2014), which then enhanced memory for conceptual information. However, providing explanations may also have drawn attention away from activities that would increase procedural knowledge, such as practising calculations.

These findings suggest that, while providing explanations can aid understanding, they also can work to draw attention away from evidence, and other relevant information (Kuhn & Katz, 2009; Williams et al., 2013). It is possible that explanation is reinforcing prior beliefs, which may negatively impact on participants' ability to evaluate new evidence (Gopnik et al., 2017; Koerber et al., 2005).

#### 1.1.3.7 Impact of socio-economic status and gender

As with the causal understanding literature, much of the research looking at scientific reasoning has been done with participants who were white, and from

middle-class backgrounds, without considering gender differences. This is the case even though both SES and gender are known to influence science attainment in school (Curran & Kellogg, 2016; Nunes, Bryant, Stran, Hillier, Rarros & Miller-Fridmann, 2017). In particular, the research looking at younger children's scientific reasoning skills, with more controlled tasks such as Blicket detector, has tended to be of this sort. It is likely that the impact of SES on causal understanding (see Wente et al., 2017, described in section 1.1.1.4) would be similar, whereby children from lower SES backgrounds would on average perform less well in these scientific reasoning tasks, compared with children from higher SES backgrounds. Other studies mentioned above, more frequently with older participants, report a broader range of ethnicity, and SES. They have not, however, explicitly assessed the impact of either on performance in those studies.

Nevertheless, there is a substantial literature looking at the impact of SES on performance in school, and on science performance more generally. For example, in a longitudinal study in the USA, Saçkes, Trundle, Bell, and O'Connell (2011) looked at the impact of children's early science experiences at age five on their science achievement at both five and eight years old, with a cohort of more than 8000 children. Their cohort was balanced across gender, and about 60% white, 10% Black or African-American, and nearly 20% Hispanic. The rest of the cohort were of less common ethnicities or mixed-race. They assessed SES using income level, and parental education. They sought information on the science classroom learning environment, the amount of science teaching, and the type of activities the children participated in. They tested the children's prior knowledge on a general knowledge test at the beginning and end of the kindergarten year at age five, as well as a third grade science achievement test at age eight. They also assessed approaches to learning including attentiveness, persistence, eagerness, independence, flexibility and coordination in learning at age five years. They found that SES, gender, prior knowledge, motivation, amount of science teaching, and engagement with science activities explained 75% of the variance in the end of kindergarten test, and that SES, gender, prior knowledge, motivation explained 61% of the variance in the third grade science achievement test at age eight. Specifically, children from higher SES backgrounds had higher prior knowledge at the beginning of kindergarten, and the impact of SES continued, although decreased, by the end of kindergarten and third

grade science achievement tests. The gender attainment gap in science was also observable when assessing prior knowledge, and appeared to increase by the end of kindergarten, and more so by the end of third grade. In contrast, girls had higher approaches to learning scores than boys. Given that approaches to learning were assessed by teachers, as was observed when discussing gender in the section on executive function (section 1.1.2), it is possible that the teachers ratings may be biased by gender, more favourable to girls, given the ratings do not appear to correspond to performance. Quinn and Cooc (2015) found that the gender gap at the end of third grade (age eight) narrowed slightly more by eighth grade (age 13), and disappeared after controlling for prior maths achievement (they used the same longitudinal study cohort that Saçkes et al. (2011) used). They also reported an ethnicity gap in science test scores, which remains constant from third to eighth grade for the Black/White comparison; decreases for the Hispanic/White comparison, and disappears for the Asian/White comparison (White students started with higher mean scores in all three comparisons).

A recent UK report (Nunes et al., 2017), focusing on the impact of SES on attainment in science, provided an extensive review of the current literature, along with analysis on the performance of disadvantaged pupils in national science tests, compared with pupils from higher SES backgrounds. The review found that the impact of SES seems to be quite robust, and has been replicated multiple times over nearly 50 years, in many different countries, with different levels of affluence, using different measures of SES and science attainment. The impact of SES can be seen at the country level, where children from low income countries appear to learn less science than those from high income countries. It is observed at the school level, where those who go to lower SES schools appear to perform less well compared with children who go to higher SES schools (measured by the mean SES of the students). There also appears to be an impact at the individual level, over and above the school level SES impact, where children from lower SES backgrounds appear to do less well than children in higher SES backgrounds in the same school. They did not find any evidence for differences in performance being mediated by differences in interest or motivation.

It has also been found that disadvantaged pupils had much lower scores in national science tests and examinations at all education levels (Key Stage 1-5, A-Level) when

compared with pupils from higher SES families (Nunes et al., 2017). The gap appears to grow over time, and seems to be largest towards the end of secondary school (age 16). The review also concluded that early general performance (at age five) is predictive of later science achievement, concurring with the findings of Saçkes et al. (2011).

Of particular interest here is the report of further analysis and data from Bryant, Nunes, Hillier, Gilroy, and Barros (2015) which looked at the relationship between SES and performance on a control of variable strategy (CVS) task, frequently used as a measure of scientific reasoning. In the original study, set in the UK, 11-year-old participants were given a CVS task, and their scores were then related to their performance in science for the following three years. The task had two components, one asked them to judge whether a particular comparison was suitable to determine the effect of the variable under question, and the second requiring a decision on the comparison based on their own opinion. Bryant et al. (2015) found that better performance on the task predicted higher attainment in science three years later, even after controlling for age and IQ. Furthermore, they found that the score when children had to compose the comparison themselves was a better predictor, than when they only had to evaluate the comparison. Further analysis of the data, showed that SES was related to performance on the CVS task, with children from lower SES backgrounds doing less well than children from higher SES backgrounds.

These findings suggest that both gender and SES has a substantive impact on science attainment. There is also some evidence to suggest that SES impacts on scientific reasoning itself.

### 1.1.4 Language

#### 1.1.4.1 Language and scientific understanding

Scientific reasoning is, to a large extent, dependent on language skills. In studies that investigate scientific reasoning, child participants must be able to understand the verbally given instructions. As such, any inferences drawn assume at least some degree of language understanding, even with younger participants. In fact, the Blicket detector paradigm was developed in part to allow testing of implicit understanding in younger participants (as young as two years of age) in causal

understanding studies (Gopnik et al., 2001), who lacked the ability to explicitly reflect on causal relations, and produce verbal responses explaining their beliefs regarding a particular causal system. Variants of the Blicket detector have frequently been used in scientific reasoning studies in younger children. This is in part because of the simplicity in operation (e.g. Cook et al., 2011; Gopnik et al., 2001; Schultz et al., 2007), which allows assessment of scientific reasoning ability in the younger population. This population would not have the ability to participate in the more complex tasks frequently used to assess scientific reasoning, such as the causal variable strategy tasks (Chen & Klahr, 1999; Dean & Kuhn, 2007), where the instructions were delivered either verbally, or in text.

Similarly, nearly everything children learn about science is mediated through language, spoken and written, either through the teacher or parents, textbooks, or communication with their peers. Furthermore, children's ability to express their understanding of science in school is largely done through the medium of reading and writing, particularly as children get older. So to effectively express their science understanding in school, children need to have adequate linguistics skills to start with. In science class, they will also have to learn the language of science; that is, to scientifically reason in an explicit way using appropriate terminology so as to question, investigate, design, test, analyse, evaluate, theorise and so on. Added to this, children are also required to learn a whole new set of concepts and labels for technical information learned about in science class. To express the new linguistic information necessary for learning, Norris and Phillips (2002) drew a distinction between fundamental and derived senses of scientific literacy, where reading and writing with scientific content is seen as fundamental, and being knowledgeable and educated in science as derived.

Much of the language of science is distinct from everyday language. Fang (2006) investigated the differences between language in science and everyday language in science textbooks aimed at middle school children aged 11 to 14 years old in the USA. They found that in science class, not only are children frequently exposed to technical words that rarely occur in everyday speech, but they also frequently have Latin and/or Greek origin and quite complex to read and understand. This is particularly the case when multiple such words are used in a single sentence. In addition to this, words sometimes mean different things in science and everyday

language (e.g. fault - break in rock formation *or* responsibility for mistake or act of wrongdoing; Fang, 2006). There are many other examples of words that are not used as they are typically used in everyday language (see Fang, 2006 for further examples); as well as other examples of differences which may lead to difficulties such as more complex sentence use, use of the passive voice, and so on. These differences place large demands on children who are learning both new vocabulary and new concepts. Children who have weaker language skills and/or are learning in their second language may be particularly likely to struggle (Kigel, McElvany, & Becker, 2015). This is notable when they go to secondary school and are exposed to much more complex language in the interaction between students and teacher or in textbooks, than in primary school.

#### 1.1.4.2 Language based factors related to science learning

Fang (2006) suggested a number of language-based interventions to overcome problems (some) children may have with linguistic aspects of science learning, such as technical vocabulary, complex words and sentences, uncommon usage of words, and so on. These were largely comprised of language-based activities, using a scientific context, which draws attention to specific aspects of scientific language that may cause difficulty for children. Fang and Wei (2010) conducted an intervention, with 10- to 11-year-olds to improve children's scientific literacy, in a school around 50% white, 35% black, with around half the student population considered to be low SES background. Students in the intervention group, received their typical science curriculum, plus reading strategy instruction, and a home science reading programme. Students in the control group received only their usual science curriculum. The study found that children who received extra reading strategy instruction, and a home science reading programme outperformed children who did not, in both a fundamental and derived sense of scientific literacy. Fang and Wei (2010) assumed that the improvements in fundamental scientific literacy could be ascribed to the reading strategy instruction, and improvements in the derived sense of scientific literacy to the home science reading programme. However, it is conceivable that reading scientific texts with parents or other family members may benefit reading and writing as well through explicit discussion of scientific topics (see below). In addition, while Fang and Wei (2010) expected there to be a benefit, establishing what aspect(s) of the intervention were doing the work is difficult in

the experimental design they used. Furthermore, it might be that students whose parents gave them permission to participate in the intervention, were parents who were predisposed towards participating in their children's learning, where if such an intervention was adopted as part of the curriculum, it would not have the same impact as not all parents or family members would participate as actively as parents or family members did here. As a result, although Fang and Wei (2010) is suggestive, it is unclear what aspect(s) of the intervention were contributing to the improved scientific reasoning scores.

There is evidence to suggest parental conversation is of benefit for science learning. For example, discussing content learned in science enhanced the memory of participants aged four- to six-years-old when asked to recall science related content discussed with parents several days later. Leichtman, Camilleri, Pillemer, Amato-Wierda, Hogan, and Dongo (2017) tested participants from two schools, both predominantly white (over 70%), one upper middle class, the other lower middle class, from the USA. Children were taught about a science topic, and then later that day parents were asked to record two conversations, one about the science lesson (they were told the topic of the science lesson, but given no other information or instructions regarding how the conversation should be conducted) and the second about a separate personal event that their child had enjoyed. They found that the conversational style of parents (e.g. open ended questions, descriptive language) predicted the amount of information provided by the child, which then appeared to have benefitted memory six days later. They did not find any differences between the schools, which is surprising given the reliability of SES differences in attainment at the school level (Bryant et al., 2017). However, one quarter (26%,16 participants) of their original sample were excluded from the final sample, the majority because their parents did not complete the parent-child conversation. The samples were also small, with only 20 participants from each school. It is possible the lack of difference between high and low SES schools is related to the dropout of participants, as opposed to there being no differences in the different samples.

The relevance of parent involvement and language has also been identified for slightly older children, in a more controlled context and with a larger sample. Philips and Tolmie (2007) examined the effect of parental support on six- to eight-year-old children's understanding of a science problem – the balance scale problem. There

were around 150 children from 10 different schools in Scotland (no information on SES was given). Participants were given the task of balancing the scale beam, where the causal variables were distance from the centre point of the scale and weight. The experimenter put an arrangement of weights on their side of the scale beam, and the participants had to balance the scale without copying what the experimenter did. They were allowed to do this with the assistance of a parent, who had received instruction in the rules guiding causal behaviour of the system. Philips and Tolmie (2007) found that participants who had parental assistance provided appropriate solutions more quickly than participants who did not, although control participants caught up by the third session. However, this was not the case for appropriate explanations, where participants with parental assistance provided better explanations in all three sessions. This benefit was most notable with children whose parents focused on verbalising the interaction between distance and weight. Furthermore, Philips and Tolmie (2007) found that it was the combination of explicit operationalisation plus high level explanation that benefited progress. So the parent providing an explanation appeared to be more effective for progress, when participants also observed the correctness of the explanation. Philips and Tolmie (2007) did not specifically look at the impact of SES here but again one would imagine that it would have an impact, where higher SES parents would be more likely to provide appropriate boosts to learning, compared with lower SES parents.

With adolescents, Gerber, Cavallo, and Marek (2001) looked at the impact of the informal learning environments on students' scientific reasoning, in 12- to 15-year-old students, where informal learning environments include activities with family, and/or friends, as well as at school. The initial sample consisted of over 1000 students, and the final sample of around 500 students, where only students with enriched and impoverished informal learning environments were included. The sample was around 80% white, and of relatively equivalent SES and academic abilities (based on teacher interviews and questionnaire data - no further information was given). They found that participants with better informal learning environments perform better on tests designed to assess scientific reasoning. One issue with this study, is that although there appeared to be a benefit of informal learning environments, the assessment of informal learning environments included activities with family, activities done alone, at school, or someplace else, and other

factors such as employment, chores and travel. It is possible that some of these activities benefit their scientific reasoning more than others (for example, parental involvement), and that examining these factors separately in the future may shed more light on what aspects of the informal learning environment are doing the work.

#### 1.1.4.3 Impact of socio-economic status and gender

With regards to SES, Bryant et al. (2017) suggested that the level of children's literacy may be a possible mediator of the relationship between science attainment and SES. Their review found that there was frequently a positive relationship between children's reading ability and their science attainment. However, they pointed out that most of the studies they reviewed assessed both reading ability and science attainment concurrently, making it challenging to establish cause and effect. They proposed a need for more longitudinal, and intervention studies. To address the longitudinal evidence, they conducted a new analysis based on the AVON longitudinal study of parents and children (ALSPAC) in the UK. This cohort included variables such as SES, IQ, reading comprehension, vocabulary, and science attainment at key stage 2 (11 years; over 5000 participants) and key stage 3 (14 years; over 3000 participants). They found that, when reading comprehension was taken into account, the relationship between SES and science attainment decreased dramatically, and even further when reading comprehension and vocabulary were both taken into account. This was also the case for scientific reasoning, but to a lesser extent. These results point to the importance of language ability in science attainment, and also suggest factors that may mediate the SES science attainment gap.

Differences related to SES in early language proficiency can be observed in children as young as one and a half to two years of age. By aged two, a six-month gap has been identified between high and low SES groups in processing skills critical to language development (Fernald, Marchman, & Wiesleder, 2013). Differences in early vocabulary development have been observed from as young as two years of age in high, compared with medium SES populations. Hoff (2003) found that higher SES children's vocabularies grew more than lower SES children's over a 10 week period. Hoff (2003) suggested that the properties of maternal speech that differed across the two groups could account for the difference. Given that the impact of SES can be seen on language from such a young age, if language ability plays a role in science

attainment, then it would not be a surprise to find an impact of SES on very early measures of science ability (see section 1.13 on scientific reasoning).

Regarding gender, in the early years of language development females tend to outperform boys. For example, girls have been found to show greater vocabulary growth over a period from around one to two-years-old (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991). Girls are also likely to out-perform boys in both specific and general measures of language up until the age of five (Bornstein, Hahn, & Haynes, 2004). Gender gaps in reading have also be observed at school, were girls frequently show better performance compared with boys (Ma, 2008). He looked at the gender gap in reading, mathematics, and scientific literacy across multiple countries, with data from the OECD's 2000 Program for International Student Assessment (PISA), based on a cohort of 15-year-old students, of around 200,000 students, from around over 7000 schools. Standardised achievement tests were administered to measure reading, mathematics, and science literacy. Ma (2008) found that there was a gender gap in reading performance that was biased towards girls in 40 out of 41 countries, with the mean difference between scores ranging from 5.31 (Indonesia) to 49.01 (New Zealand). The extent of the difference ranged across countries. Most countries, including the UK, showed only small female advantage. This is in contrast to mathematics where 29 of 41 countries showed a performance in favour of boys, with the mean difference between scores ranging from 6.15 (UK) to 25.07 (France). The UK showed a small male advantage. For science, in 14 of 41 countries there was a performance bias towards boys, with mean difference scores ranging from 7.12 (Mexico) to 23.66 (Poland); and 5 of 41 countries showed a performance bias towards girls, with mean difference scores ranging from 5.50 (Thailand) to 19.73 (New Zealand), where the rest showed no difference. Interestingly, all five of the countries that show a female science bias, also show a relatively large female reading bias.

The trend towards better male performance in science attainment (see section 1.1.3. on scientific reasoning), even though they are on average likely to do less well than girls in tests of reading, suggests that performance in science attainment cannot be purely explained by linguistic factors. However, they do appear to play a role, and therefore should be considered, both when attempting to gain an understanding of the development of scientific reasoning, as well as when considering a science

curriculum. This is especially the case if one wants to improve the performance of children from lower SES environments in science.

In conclusion, language is likely to be relevant to scientific reasoning both on fundamental and derived levels, such that science learners with poorer language ability are likely to demonstrate poorer performance.

### 1.1.5 Science education

The goal of the national curriculum in the UK, regarding science education, is to "provide the foundations for understanding the world to the specific disciplines of biology, chemistry and physics…and all pupils should be taught essential aspects of the knowledge, methods, processes and uses of science" (Department for Education, 2015, para. 1).

At the beginning of primary school, science education is largely exploratory, and seeks to encourage children to both become interested in the world around them, and to use different types of scientific enquiry to answer their own questions. As primary school continues, children are expected to start making decisions about what types of scientific enquiry may be best to answer questions they have about everyday phenomena. By the end of primary school, their knowledge and understanding are expected to have become more abstract, enabling them to be able to recognise how these more abstract ideas help them to understand and predict how the world works. They are expected to be able to answer science questions using different types of science enquiry such as observation over time, looking for patterns, carrying out fair tests, and using a wide range of secondary sources of information. By age 16 (the end of the prescribed national curriculum), children are expected to have a good understanding of the subject disciplines of biology, chemistry and physics, whilst also understanding the role that science plays in our lives. Scientific thinking is expected to underlie their learning throughout. It could be said that teaching scientific reasoning is one of the primary goal of science education, given the large majority of children leaving school do not go on to use the specific information regarding physics, chemistry, or biology (possibly except for information relating to human biology and development). The national curriculum for science (in the UK) aims illustrate this, indicating that their aims are to ensure that all pupils:

"develop scientific knowledge and conceptual understanding through specific disciplines of biology, chemistry and physics; develop understanding of the nature, processes and methods of science through different types of science enquiries that help them to answer scientific questions about the world around them; are equipped with the scientific knowledge required to understand the uses and implications of science today and for the future" (Department for education, 2015, para 2)

The latter two of the three aims of the national curriculum directly relate to processes typically used in scientific reasoning.

However, Li and Klahr (2006) claimed that viewing scientific reasoning as invoking an unordered set of relevant skills is problematic and does not necessarily lead to scientific reasoning. They suggested that it should be taught as a set of inter-related problem-solving strategies, and have provided suggestions for how one might implement that in the classroom. Of course, explicitly teaching scientific reasoning to children requires the teachers to have a deeper explicit understanding of the mechanics of scientific reasoning, and that the education system in place facilitates a scientific education that does more than tick the scientific reasoning checklist.

However, national curricula have a tendency to generate lists of methods, processes, skills, and content the students should gain understanding of through the teaching of science, for example. The current national curriculum for science in the UK provides examples of this at each educational stage (Department for Education, 2015, Key stage 1 programme of study – years 1 and 2). For example, during the key stage 1 programme study (five- to seven-year-olds), children should be taught to ask simple questions, and recognise that they can be answered in different ways; observe closely using simple equipment; perform simple tests; identify and classify; use observation and ideas to answer questions; and gather and record data to help answer questions. By key stage 4 (14- to 16-years-old), students are expected to show understanding and first-hand experience of the development of scientific thinking; experimental skills and strategies; analysis and evaluation; and vocabulary, units, symbols and nomenclature (Department for Education, 2015, Key stage 4).

In conclusion, scientific reasoning skills can be seen from around four years of age, with continued development throughout the school period. This development is

slow, and requires extensive educational support (Morris et al., 2012), and appropriate epistemological understanding takes a long time to develop to its highest level, which may not be attained by all individuals. This suggests that the acquisition of scientific reasoning skills require more formal teaching compared to the acquisition of causal reasoning skills, which function implicitly and are evident from a very young age.

## 1.2 SOURCE RELIABILITY

### 1.2.1 DEFINITION

One factor that is particularly relevant for scientific reasoning is source reliability. Both children and adults are faced with a plethora of information in their day to day lives that they are expected to use to make judgments about how they should act. This information comes from a number of different media (more than ever before) - such as newspapers, television, the Internet, advertisements, politicians, doctors, scientists, and scientific journal articles (where with the advent of the World Wide Web, the lay population has direct access to evidence reported *by* the scientist). When learning about science, children also receive information from a wide variety of sources, both in school and out. For example, students receive information from teachers, their peers, text books, the Internet, their parents, and so on. These different sources are differentially reliable. That is, some sources of information are more likely to provide correct information than others. For example, a doctor who has received several years of education relating to the health of human beings is expected to be more likely to be reliable with regards to information regarding health (their area of expertise) than information based on the personal opinion of someone writing on a blog (on a topic that is not their area of expertise). For a school child doing a test in science, trusting the information provided by the teacher who prepared and is guiding the lesson, is likely to be more sensible than trusting information provided by one of their peers, if that information is in conflict.

Consequently, it is important that by adulthood people are able to appropriately evaluate the reliability of sources, in order that they make good decisions that give the best possible outcomes in their lives. Given the crucial importance of source reliability, gaining a greater understanding of the role of source reliability in

children's reasoning about their everyday world is desirable, so that it can be incorporated into science education.

### 1.2.2 Children's understanding of source reliability

As stated, children are faced with information from multiple sources that they can incorporate into their reasoning about the world. Given that at a young age they are not capable of obtaining anything but the most obvious (observable) information for themselves, they rely on the testimony of their parents, teachers, peers, and, increasingly, the media where many children under the age of three now use screen based media (Duch, Fisher, Ensari, & Harrington, 2013). The *selective trust paradigm*, developed by Harris and colleagues (e.g. Koenig, Clément & Harris, 2004; Koenig & Harris, 2005), was designed to look at whether children as young as three or four years of age were sensitive to the reliability of the source of the information, and whether source reliability affects their understanding of outcomes.

#### 1.2.2.1 Research using the selective trust paradigm

In the typical selective trust paradigm children are presented with an accurate, and an inaccurate informant, usually by watching a video of the informants. These informants are identified as such during the *familiarization phase* in the following way. Children are shown a known object, such as a ball. The *accurate informant* consistently labels the known objects *correctly.* For example, when shown a ball, they say "that is a ball". The *inaccurate informant* consistently labels the known objects *incorrectly*. For example, when shown a ball, they say "that is a shoe". After the reliability of the informants had been established, children enter the test phase, where they witness the informants label unfamiliar objects. They then participate in a number of trials - an explicit reliability trial where they are asked questions such as "Did any of them [the informants] say something right/wrong?" Following that, children are asked to identify an unfamiliar object which had been identified differently by both the accurate and inaccurate informant. For example, when the novel object has been revealed, the accurate informant identifies it as a 'mido' and the inaccurate informant as a 'toma'. The experimenter then asks the child if the unfamiliar object is a 'mido' or a 'toma'. Finally, children receive a second explicit reliability trial, where they are asked "One of these people kept saying something right/wrong. Which one kept saying something right/wrong?" (Koenig et al., 2004).

Using the selective trust paradigm, Koenig et al. (2004) investigated three- and four-year-old participants' ability to discriminate between accurate and inaccurate informants. The participants were from university-based childcare centres (at a prestigious university) in the USA, with equal genders. They found that children appeared to be able to differentiate between the two informants, when asked if they had done something right/wrong at both time points. They also found that those participants who *were* able to discriminate between the two informants (around 50% of three-year-olds, and 70% of four-year-olds) were above chance (around 65%) at keeping track of an informant's previous accuracy, able to use that information to judge whether an informant should be trusted upon receipt of new information. They did not find an effect of age, although the four-year-olds appeared to be more competent than the three-year-olds. That is, from a young age, some children appear to be capable of evaluating source reliability and changing their behaviour accordingly.

This understanding of source reliability appears to be quite sophisticated. For example, Koenig and Harris (2005) recruited children from both the university childcare centre, as well as the local Head Start centre (Head Start promotes school readiness of children under five from low income families; U.S. Department of Health & Human Services, Office of Head Start, n.d.) in the USA. They do not report what proportion from each childcare centre. Approximately 60% of the participants were white, and the rest different ethnicities. Most were from primarily middle class backgrounds, and the sample was balanced by gender. They used a similar paradigm to Koenig et al. (2004), except that with the explicit reliability questions, they were asked which of the informants was "not good at answering questions". They found a similar pattern of results to Koenig et al. (2004) with regards to explicit reliability questions, and ignorance was a favoured explanation. However, they did find an age difference, where the four-year-olds exceeded chance in using information from the accurate informant to endorse claims and predict future assertions, whereas three-year-olds did not. In a second experiment, participants were presented with accurate versus ignorant informants, finding that both age groups showed a preference for accurate informants. A third experiment assessed whether children would show preferences for more reliable informants in domains beyond where they had observed differential reliability (a greater proportion of these children

were white (80%), compared with the other two experiments, and were described as coming from middle to upper class backgrounds). They found that participants preferred to ask questions of the accurate informant, with regards to the same and different domains. They do not find an age difference. Koenig and Harris (2005) did not report any performance differences dependent on type of childcare centre. This would be of interest given that Head Start is designed to provide early years resources for low SES families. As discussed in each of the previous sections, low SES can impact on the development of cognition in many different domains. Although the participants are reported as coming from middle, and middle to upper class backgrounds, it is unclear how this tallies with participants recruited from the Head Start childcare centres. It is possible that age differences, or lack thereof (in Koenig et al., 2004 as well), reflects sample variation in age groups due to participants coming from different SES backgrounds. It would have been interesting to know if there are any performance differences between childcare centres. One would predict that the children from the University childcare centres perform better than those from the Head Start childcare centres.

There is now a substantial body of literature using paradigms similar to that of the selective trust paradigm outlined above that suggests that children as young as three- and four-years-old can discriminate between more and less reliable sources. For example, Scofield and Behrend (2008) manipulated reliability of the informants, *after* informants had provided information that the participants had to decide to endorse or not. They tested three- and four-year-old participants from white middle class families in the USA. They found that just over 50% of the four-year-old participants, and just over 25% of the three-year-old participants reversed their trust in the face of newly discovered unreliability of an informant.

Pasquini, Corriveau, Koenig and Harris (2007) recruited three- and four-year-old participants from a childcare centre in the USA serving a broad SES range, of which 50% were white, and the remainder a range of ethnicities. They sought to look at the relationship between false belief and selective trust. A potential explanation for three-year-olds underperforming in the selective trust studies is that they have difficulty interpreting false labels because they do not yet understand the false belief that may motivate them. In the selective trust task not only did the researchers manipulate accuracy, but they also manipulated relative accuracy, where

informants could be accurate 100%, 75%, 25%, or 0% of the time. They compared 100% and 0%, 100% and 25%, and 75% and 0%. They found little evidence of a relationship between false belief understanding and success in the selective trust task, where children's understanding of false belief did not predict performance, even after controlling for age. They also found that four-year-olds were above chance in all three comparisons whereas three-year-olds were above chance for 100% and 0%, 100% and 25%, but at chance for 75% and 0%. That is, three-year-olds seemed relatively unforgiving of the errors made by the source who was correct 75% of the time; whereas four-year-olds accepted that the source would be more generally reliable. This age difference in performance was replicated in a second study, which also compared 75% and 25%.

When reliability was manipulated by comparing child and adult informants, three- and four-year-old participants, from the USA (no more information given) doing a selective trust task, preferred information from the adult informant when both were reliable, and when the adult informant was reliable and the child informant not reliable. However, even though they showed a strong preference for adult informants, when the child was reliable and the adult unreliable, they showed a preference for the reliable child (Jaswal & Neely, 2006).

Children also appear to prefer a consensus among sources. Corriveau, Fusaro and Harris (2009), doing a selective trust task with three- and four-year-old children (from preschools near the University, mostly white, with a range of ethnicities and SES represented - no proportions were given, from the USA). Children witnessed four informants, where there was a consensus among three of them. In both age groups children tended to accept information made by the majority. Furthermore, when faced with a choice between one of the majority, and the dissenting informant, they preferred the informant from the majority group.

The majority of this research has been done with children aged around three to four years of age, with participants that come from middle-class families. Although some participant samples included children from lower SES environments, they did not assess the impact of SES. One study (Koenig & Woodward, 2010) involved two-year-old participants who were recruited via advertisements or mailings, and were approximately 50% white, interacting with an accurate or an inaccurate informant

(the experimenter), observing how the children responded to information that came from an inaccurate or accurate informant. It was found that children responded more systematically to information from the accurate informant. They did not find an effect of gender. Of interest here however is that they recorded a measure of vocabulary levels, which they used to form two groups, one high and one low vocabulary group. They found that participants in the high vocabulary group showed more sensitivity to inaccurate sources compared with participants in the low vocabulary group. Given that poorer language skill is more likely for children coming from lower SES environments (see section 1.1.1 on language), it is possible that this is initial evidence of the impact of SES on selective trust. The majority of the selective trust literature mentioned previously does not specifically address the impact of SES or gender on source reliability evaluation.

#### 1.2.2.2 Epistemic understanding of source reliability in young children

The literature is extensive and children show quite sophisticated preferences and background knowledge regarding who is likely to be reliable or not. For example, they appear to understand that an informant who is accurate in one domain is not necessarily accurate in other domains. However, they are also likely to believe that an inaccurate informant in one domain should be avoided in another (Koenig & Jaswal, 2011). Selective trust has also been shown across domains, in contexts other than word learning. For example, it can be seen when children are learning new object functions (Birch, Vautheir & Bloom, 2008); finding a target (Nurmsoo & Robinson, 2009a) or deciding whose advice to accept (Vanderbilt, Liu, & Heyman, 2011). Children may also discriminate depending on how accurate sources have achieved their prior accuracy. Four- to five-year-olds preferred sources who did not rely on help from a third party (Vanderbilt et al., 2011).

The fact that children are likely to show a preference for reliable sources of information in many different contexts has led to the claim that children are showing epistemic awareness regarding the knowledge of the informants (Koneig & Harris, 2007). An alternative conception is that children base their responses purely on the output of the informant, without making any inferences regarding the informant's interior knowledge. This alternative view suggests that the oddness of the behaviour of the inaccurate informant mis-naming common everyday items may be enough to

explain the preference for the accurate informant, without the need to infer the existence of epistemic awareness.

Lucas and Lewis (2010) challenged Koenig and Harris' (2007) claim that children are showing epistemic awareness. They suggested firstly, that if children really had epistemic awareness, then they would have an understanding of misinformed knowledge. That is, they would be able to distinguish between informants who had a good reason for inaccuracy (lack of exposure to conventional information or a lack of expertise), and those who were inaccurate for no apparent reason. Secondly, Lucas and Lewis (2010) suggested that children should understand the nature of misinformed knowledge whereby they can predict and forgive misinformed information.

Mixed findings have been found when looking at children's understanding of misinformed knowledge. Nurmsoo and Robinson (2009a), using puppets and a specific object finding task, found that children could distinguish between informants who had reason to be inaccurate and those who did not. They involved four- and five-year-old participants from white working and middle-class areas in the United Kingdom. The study differed slightly from the standard selective trust task, in that the source reliability manipulation was face-to-face. They decided to use a puppet to avoid having an apparently fully formed adult giving obviously inaccurate information to the child participant. In this study, based on the format of the selective trust paradigm, reliability of the informants was established, with two completely unreliable informants. However, the participants were aware that one unreliable informant was uninformed regarding the identification of the target toys (unreliable without evidence), whereas the other informant was informed (unreliable with evidence). Participants were then asked about a target toy in test trials. Once the participant had identified which toy they thought it was, the informant puppet then (always) contradicted them, and the child was asked if they want to switch to agree with the informant puppet. The participants were much more likely to switch if the information came from the uninformed informant (unreliable without evidence; about 70% of children switched). There were no age differences. At first glance, this appears to be support for epistemic awareness regarding the knowledge of informants, whereby children appear to predict and forgive misinformed information.

However, in a second experiment, using video informants and a generalisable word learning task, children as old as seven years of age failed to distinguish between the two types of inaccurate informants. Nurmsoo and Robinson (2009b) used four-, five-, and seven-year-old participants from white working and middle-class areas in the United Kingdom. They had two groups of participants who did slightly different tasks. One task was similar to the standard selective trust task. The second task was a setup similarly to Nurmsoo and Robinson (2009a), in that they had two inaccurate informants, providing word labels for objects. One inaccurate informant was blindfolded (unreliable without seeing object) and the second inaccurate informant could see (unreliable with seeing object). Once source reliability had been established in the test trials, both informants could see (i.e. the blindfold was removed), and the participant was asked which word label they preferred for new objects that had just been identified by both informants. In the first group, as has been previously found, children were sensitive to informant accuracy, and preferred to learn from accurate informants. However, for the second group, participants failed to discriminate between reasons for inaccuracy in the two informants, solely basing their responses on informants' history. This was even though they seemed to understand that the blindfold could affect familiar object naming. Nurmsoo and Robinson (2009b) replicated the result, adding in previous reliability information that the blindfolded informant was accurate when not wearing the blindfold. Whereas Nurmsoo and Robinson (2009a) found children did appear to pay attention to whether the informant was misinformed, Nurmsoo and Robinson (2009b) found that they did not appear to do so, even at the age of seven.

One likely explanation for these puzzling results might be that communicators and social cues necessary to engage in mentalistic reasoning (thus demonstrating epistemic awareness) may be missing from the task using video informants. None of the usual non-verbal information usually available to direct children's inferences regarding internal knowledge was present, and children are sensitive to non-verbal information. For instance, children have been found to prefer informants who receive bystander assent such as nods and smiles, versus bystander dissent such as head shakes and frowns (Fusaro & Harris, 2008). However, Nurmsoo and Robinson (2009a) used puppets, and it seems unlikely that puppet informants provided more informative communicatons and social cues than a video of a human informant. A

more compelling explanation for the conflicting findings might be that children are particularly intolerant of inaccurate informants when they are learning generalisable information (new words) compared with learning specific information (identity of a hidden toy). It matters if one uses a word label incorrectly, so the nature of the inaccurate informant is important. It does not matter if the toy is not what the informant says it is. Clearly context is important, and it may be that children take source reliability into account when they deem the outcome important enough for it to be considered. This may be the case for the younger children in the study by Bernard, Proust and Clément (2015). They found that younger children (four- to five-year-olds) preferred consensus over reliability, whereas older children display the reverse pattern (six-year-olds).

#### 1.2.2.3 Other research on source reliability understanding

There has been little research looking at children's understanding of source reliability that does not use some variant of the selective trust paradigm. One such example, Fitneva (2001, see also 2008) looked at source reliability from a linguistic perspective. Fitneva was interested in how children use epistemic information in their judgments, and in the distinction between the source of information and speaker attitude (such as degree of commitment), when considering the reliability of statements. The prevailing view was that reliability of information from a speaker was judged by 'speaker attitude' and that 'source of information' contributed to the evaluation of speaker attitude (along with other epistemic devices such as lexical, intonational, or grammatical, which can characterise the origin, nature, and limit of the knowledge expressed by the speaker). However, Fitneva (2001) claimed that both elements were important when making judgements in the everyday world. Speaker attitude and source of information do not necessarily provide concurrent information. For example, a speaker may have high degree of commitment, but may not be a good source of information regarding an event (compare, for example, someone who witnessed an event versus someone who heard about an event from someone else; the direct witness would be considered to be a better source of information than the person who heard about it from someone else). Fitneva argued that both convey information regarding the reliability of information, but one would use them in different situations. Speaker attitude is used when the speaker is capable of competently deciding on the reliability of the information, and source of

information when the reliability of the source of information is questionable, whereby the speakers could debate the relevance of the information.

In an experiment that draws on grammatical pointers in Bulgarian that differentiate between events that have been reported to the speaker (first-hand information), and events that have been inferred by the speaker (second-hand information), Fitneva (2001) argued that this difference maps neatly on to the difference between source of information and speaker attitude interpretations. This allowed Fitneva to evaluate whether six- and nine-year-old Bulgarian participants (no other information was given) differentially used source of information and speaker attitude when evaluating reliability of information. Were they to, it would suggest that the source of information is directly relevant to the evaluation of the reliability of statements. The task involved participants hearing a story and being asked which protagonist they believed. The first set of stories involved searching for a location, where one would expect first-hand information (source of information) would be more useful than second hand information (speaker attitude). She found that nine-year-old participants were more likely to believe the informant that provided first-hand information, when location was important, and did not show a preference for first over second-hand informants when location was not (the six-year-old children were also showing a trend in that direction). In a second study, Fitneva (2008) found that both six- and nine-year-old Bulgarian participants, from middle-class neighbourhoods were sensitive to modality information when making reliability judgements. Participants were asked to choose between information from two different sources, where perceptual (direct perception versus hearsay) and cognitive (direct inference versus report of inference) domains were manipulated. It was concluded that older children prefer perceptual sources (i.e., sources that claim to have observed the event), whereas younger children prefer cognitive sources (i.e., sources that claim to know). Older children were also thought to prefer first-hand information, whereas younger children did not appear to discriminate between first and second hand information.

These results suggest that older children (and some younger children) might be using epistemic information to evaluate the reliability of statements, where they distinguish between statements based on minor grammatical changes. However, given that the effect was not strong in six-year-olds, it seems unlikely that three to

four-year-olds would access the epistemic information in the same way. Fitneva's (2001, 2008) results have provided a further challenge to the claim that three to four-year-old children are accessing epistemic information when they demonstrate a preference for the accurate informant.

Nurmsoo and Robinson's (2009 a & b) contradictory findings may be offered a resolution here. In Nurmsoo and Robinson's (2009a) study the children were asked to make a perceptual judgment regarding an object. As such, the reliability of source of information was important enough for them to pay particular attention to what the source of information may have perceived. However, in Nurmsoo and Robinson's (2009b) study children were asked which label they prefer from the 'reason for inaccuracy' and 'no reason for inaccuracy' informants. As the cognitive process by which a naming error occurs cannot be observed, the source of information may have been less salient to the children during the training phase (compared with when the informant is asked to make a perceptual judgment). When they were asked which label they preferred in the test phase, that lack of salience led to a failure to discriminate between informants who had good reasons for inaccuracy and informants that did not. In the standard selective trust tasks, the fact that an informant is inaccurate (compared with an accurate informant) is salient enough for the children to discriminate between informants. Furthermore, in another study requiring children to use perceptual information (identify the contents of boxes), Mills, Legare, Grant and Landrum (2011) found that in some cases, three- to five-year-old children could distinguish between accurate, ignorant (self -identified) and inaccurate informants. They were better at doing this when the ignorant informants explicitly stated their lack of certainty, a clear indicator of the reliability of the source of information. However, they were not always able to do this, and when the task was more epistemically challenging (fewer overt cues), they often failed. A similar pattern was found by Vanderbilt et al. (2011), in another location task. The sample consisted of three-, four- and five-year-olds, 50% of whom were white, from the USA (no information on SES was given). Source reliability was established where children were shown a video of an adult informant who either helped or tricked other actors when trying to locate an object. Participants were then asked to locate an object. Five-year-old children preferred informants who were 'helpers' over 'trickers'. By contrast, three-year-old children appeared not to

recognise that deception had occurred. Four-year-old participants did appear to recognise the deception but it appeared not to influence their preference. Vanderbilt et al. (2011) suggested that the four-year-olds failed to understand the implications of their knowledge regarding helpers and trickers, which may explain the mismatch between knowledge and behaviour. This is in keeping with the idea that younger participants lack full epistemic awareness regarding what informants might know.

Given this background, the claim that children as young as three years of age are demonstrating fully specified epistemic awareness regarding what the source knows seems unlikely. If Lucas and Lewis's (2010) criteria are addressed - looking at what children understand when faced with misinformed knowledge - children frequently fail to discriminate between more and less reliable sources.

#### 1.2.2.4 Prior beliefs and knowledge

Another problem with the research looking at children's understanding of the role of source reliability is that the paradigms use information from differentially reliable sources on topics on which children are unlikely to have strong beliefs. Reliability of the sources is often manipulated externally, whereby children witness the source as being more or less reliable, or having more or less expertise. The information they receive does not contradict what they already know, or if it does (e.g. mislabelling objects), it is done by the unreliable source. It is unclear what would happen when children are provided information that is contradictory to what they know. In the literature on scientific reasoning (see section 1.1.3), prior beliefs have been shown to affect children's scientific reasoning (e.g. Gopnik et al., 2017; Koerber et al., 2005), and that this effect increases with age (e.g. Gopnik et al., 2017).

In science class children are frequently provided with information from nominally reliable sources, teachers, which may differ from their naive beliefs, regarding physical objects for example. How they incorporate this information into their reasoning, and how the source of this information impacts on that is important to understand given the ongoing desire for children to leave school as efficient scientific reasoners. Landrum, Eaves Jr, and Shafto (2015) suggested that learning from other people, which includes an appreciation of source reliability, requires the integration of reasoning about an informant's psychological properties, and reasoning about the implications of the information presented by the informant.

Predictions include a preference for more reliable sources, as outlined earlier. It is also predicted that learners recognise the importance of both an informant's knowledge *and* intention. That five-year-olds more often prefer helpers with good intentions (Vanderbilt et al., 2011), and nice experts rather than mean ones (Landrum, Mills, & Johnston, 2013) provides evidence for this contention. Landrum et al. (2013) found that predominantly white middle class participants were more likely to prefer information from previously established experts, and that this increased with age. When considering whether expertise in competition with benevolence would prevail, children were provided with information regarding informants' expertise, and whether they engaged in nice or mean behaviour. They found that information regarding niceness/meanness was more influential than information regarding expertise. They were likely to prefer the nice informant, even if they were a (mean) expert (Landrum et al., 2013). It is possible that what children perceive the informant as intending may dominate preference, particularly at an early age, and may be why they prefer cognitive over perceptual modalities (Fitneva, 2008). Adults also pay attention to the motivation of those providing information (Kunda, 1990), including believing that others are more susceptible to manipulation than themselves (Pronin, Gilovich & Ross, 2004).

When looking at what factors older children consider when judging the reliability of sources, Durkin and Shafto (2016), used a format similar to the selective trust paradigm, in an academic domain. They found that source reliability also affected learning a mathematical topic in nine- to 11-year-old participants, nearly all white, from middle and upper-class backgrounds. Reliability of the source training was established via worked examples, whereby two informants provided written examples that were either always or never correct, and two informants that were both inconsistently correct 50% of the time. Participants were assessed on their knowledge regarding the topic before and after the training. They showed that receiving information from the high reliability informant, who provided only correct examples, improved the learning in the nine-year-old participants, who had no instruction in the mathematical topic. 11-year-old participants, who already had some experience in the mathematical topic, showed more learning with the low reliability informant, who provided both correct and incorrect examples. This is possibly because older participants had to reinforce their learning by considering

whether the example was correct or not, which they did not have to do with high reliable informant. However younger participants, having no experience with the topic, found accurate examples more illuminating, compared with examples that had 50% chance of error.

Another assessment of source reliability understanding in an academic environment was done with 11- to 13-year-olds, from an inner-city school in the USA, which was predominantly African-American (60%), and low SES (60%) with 40% not meeting reading ability levels expected for their age. These participants were given a question and asked to rate the usefulness of a number of different sources which were characterised by six attributes (title, author, where published (e.g. newspaper or textbook), type of source (e.g. letter, editorial), date published, and brief summary). Each of the six attributes was varied according to what one might come across during an Internet search. Experts designated the set of sources as useful, or not useful based on the question. They had to rate all six attributes as well as providing a holistic rating on a three-point scale ranging from useful to not useful. Braasch et al. (2009) wanted to investigate whether students who were better and poorer at differentiating useful/less useful sources based their judgements on different types of attributes. They divided their participant group into three based on performance, and used the top and bottom performing group for the analysis. They found that, for participants who were better at differentiating between the usefulness of sources, the only attribute that correlated with holistic rating for *useful* sources was summary, whereas for *not useful* sources, holistic rating was related to both summary and title. That is, they appeared to pay attention to content information when making usefulness judgements. However, for participants who were less good at differentiating, for *useful* sources the only significant correlation was between author and holistic rating, and for not useful sources the significant correlations were related to author, and venue of publication, and holistic rating. It appears that these participants were paying less attention to source content when making their judgements. One might expect that where a source was published would play a role in its evaluation, however it does not appear to with either group of participants. But the sample size was very small (33 participants in the analysis), and spurious significant correlations are common among multiple comparisons. Even though there appeared to be differences in

attention to attributes between people who were good at differentiating, and those who were not, one would be reluctant to draw any strong conclusions until this result is replicated. The sample is also largely students from a lower SES background. Given that these participants are known to perform less well in reading (see section 1.1.4 on language) it may be that performance in participants from a higher SES background would show a different result. For example, holistic ratings of each source may correlate with more than one attribute, where published and type of source also give an indication of reliability, as well as summary.

In slightly older middle class Norwegian children, Bråten, Ferguson, Strømsø, and Anmarkrud (2012) investigated what type of justification 14- to 15-year-olds preferred in relation to knowledge claims in science. That is, how did they rate different types of sources. They compared personal justification with justification by authority (information based on scientific research and conveyed by teachers, textbooks, and scientists), and justification by multiple sources. Participants filled in a justification for knowing questionnaire, as well as completing three short essay questions on a scientific issue such as explaining the relationship between sun exposure, health, and illness. Participants were rated on how well they explained the issue, and integrated different perspectives discussed in the source documents. They found that the teenagers preferred justification by authority, followed by multiple sources, then by personal justification. They also found that a preference for personal justification negatively predicted good performance on the essays, and that justification by multiple sources positively predicted good performance on the essays. It is possible that these tasks discriminate between higher performing children who have achieved an evaluativist level of epistemological understanding, as compared with the low performing children who have only reached a multiplist level of epistemological understanding (Kuhn, 2010). It is also worth noting that the higher performing participants appear to be paying attention to the same type of evidence that adults pay attention to, source reliability (justification by authority) and strength of evidence (justification by multiple sources; Corner & Hahn, 2009).

There appears to be little direct research in the developmental literature that specifically looks at the impact of SES or gender on source reliability understanding. When gender is mentioned, is largely to state that preliminary analysis has not found an effect of gender, which is therefore not included in the main analysis. Given that

language plays a role in evaluation of sources with older children, where the tasks tend to utilise literacy based tasks, it is possible that there would be an impact of SES.

### 1.2.3 ADULTS' UNDERSTANDING OF SOURCE RELIABILITY

Adults also take source reliability into account. For example, research that takes a Bayesian approach has found that people rate arguments from more reliable sources as being more convincing (e.g. Hahn, Oaksford, & Bayindir, 2005; Hahn, Harris, & Corner, 2009). Here, source reliability is manipulated in a much more naturalistic way. People are asked to evaluate information from sources they are likely to have come across in their everyday lives (such as a research body vs. TV interview in Hahn et al., (2005); or information that comes from journal article vs. an advertisement in Hahn et al., 2009).

Adults also show a wishful thinking bias (Gordon, Franklin, & Beck, 2005). This concurs with research that children appear to reflect on the intention of the source of the information, such as preferring kind over mean sources, to the extent that they prefer the kind low reliability source to the mean high reliable one (Landrum et al., 2013). Gordon et al. (2005) found that participants showed a bias, in that they thought to attribute desirable predictions to the reliable source and vice versa. This effect was observed whether source reliability information was available at encoding, or only at retrieval.

Similar phenomena have also been seen in research that looks at the role of politics as it affects decision relevant science (see the work by Dan Kahan), where people's beliefs about climate change, for example, tended to follow their political leanings. In this case, people appeared to believe sources that cohered with their own (politically based) thinking, without placing as much importance on actual source reliability. Contrary to popular belief, Kahan, Peters, Dawson and Slovic (2014) found that science comprehension did not reduce the effect. Instead they found that a higher ability and disposition to make use of quantitative information increased, rather than decreased, polarisation. It is possible that these errors in evaluating source reliability in these particular circumstances are to do with the assessment of the motivation of the source, where people downplay the information because they believe the source has ulterior motives.

In conclusion, both children and adults have been shown to pay attention to source reliability. However, it is unclear at what age children begin to show epistemic understanding regarding what the differentially reliable sources might know. Furthermore, it is not clear how well these findings regarding children's understanding of source reliability explain how children understand and use source reliability information in their everyday environments.

## 1.3 AIMS OF THESIS (RATIONALE)

The aims of this thesis were to assess the development of understanding regarding the role played by source reliability when reasoning about causal systems. Understanding source reliability is a crucial aspect of scientific reasoning, and is particularly important in the 21st-century, when, with the development of the Internet, there is an extraordinary range of sources pertaining to a particular (scientific) topic. Although there is a wide body of research focusing on source reliability in younger children, this largely involves paradigms such as the selective trust paradigm (see section 1.2.1), where source reliability was frequently artificially manipulated (for example, the child would observe one proponent in the experiment make more errors than the other, be they human collaborators, video actors, puppets, or other sources used in these experiments).

As a result of children's appropriate responses in these source reliability tasks, it has been concluded that even three- and four-year-olds are capable of incorporating information regarding source reliability into their reasoning about the world around them (e.g. Koenig & Harris, 2005). It is also claimed that they show epistemic awareness regarding the knowledge of the informants (Koenig & Harris, 2007). However, Nurmsoo and Robinson (2009b), looking at children's understanding of misinformation, did not find an understanding of source reliability even in seven-year-olds. Furthermore, research using a different paradigm altogether suggested that nine-year-olds paid attention to source reliability, but six-year-olds did not (Fitneva, 2001), indicating a disparity in the age at which source reliability understanding is observable in children. As such, it is unclear at what age children show epistemic awareness regarding the knowledge of the informants and this needs further clarification.

One way to add clarification is to use research methods that more closely reflect children's day to day experiences. Research that focusses on source reliability understanding in more naturalistic environments, using more everyday activities, is less frequent. The selective trust tasks tend to rely on an artificial manipulation of source reliability, where children may have few preconceived conceptions regarding the reliability of the informants at the beginning of the task, which is unlikely to be the case in real life. When a more naturalistic paradigm was used (Fitneva, 2001), the age at which children appeared to discriminate between sources was much older (nine years old) than is typically found in selective trust tasks (three- to four-year-olds). It may be that, whilst children can discriminate between artificially manipulated sources where the differences between them are obvious, in real life they may find it much more difficult, such that appropriate discrimination between sources would not be observed until children are older. Furthermore, even if children are able to discriminate between clearly defined sources, it is not clear that they would be able to use that information to guide reasoning in more naturalistic environments.

There is research using more naturalistic contexts, but this has focussed on older children and adolescents, for whom there is less evidence regarding the relevance of source reliability. Furthermore, the naturalistic paradigms used to investigate older children and adults' understanding of source reliability tend to be very different from those used for younger children. They often manipulate source reliability in more naturalistic ways (e.g. participants are required to evaluate actual sources, Braasch et al., 2009; or consider scenarios that potentially could come from real life, Hahn et al., 2005), and frequently are embedded in academic environments (Braasch et al., 2009). However, these tasks are too difficult for younger children to do, so preclude direct comparison of performance.

It is likely that the development of implicit and explicit understanding of causal systems take place at different rates. For example, as the literature on the development of causal understanding (see section 1.1.1) suggests, children's intuitive understanding of simple causal systems appears to develop at a very young age (e.g. Schulz & Gopnik, 2004). Examples of discrimination between sources also appears to be evident from a young age (e.g. Koenig & Harris, 2005). However, explicit understanding is likely to appear later during the primary school years into

adolescence (e.g. Klahr et al., 1993), with input from science education in school. These two development trajectories are rarely compared. Using a paradigm with participants across multiple age groups would allow the developmental trajectories to be compared.

Although source reliability has been established as an important component when reasoning scientifically, and adolescent and adult tasks frequently assess source reliability within a scientific context (e.g. Bråten et al., 2012; Hahn et al., 2005), this is not the case for the literature looking at younger children's understanding of source reliability. It is important to understand the development of components of scientific reasoning, such as source reliability, for the purposes of developing the most effective strategies for teaching science, and scientific reasoning. Furthermore, it would be useful to provide a point of comparison between the adult literature, which tends to focus on ratings for the strength of arguments in source reliability, and the child literature which usually focuses on discrimination and endorsement in younger children, and evaluation of sources in adolescents.

As such, the aims of the thesis were firstly, to investigate the development of epistemic awareness in relation to what sources might be assumed to know. This was done by manipulating source reliability more naturalistically, and examining how it is related to reasoning regarding a familiar causal system in a familiar environment – school. The paradigm used, consequently, incorporated activities that were not substantially different from activities that might be learned about in a science class, where scientific knowledge is gained both through receiving knowledge from a source (such as a science teacher), and through direct enquiry and experimentation. To do this, participants were given unexpected information regarding a familiar causal system, from a more and less reliable source, or were given no information.

Secondly, in order to gain a greater understanding of the developmental trajectory, implicit and explicit understanding of a specific causal system was investigated in both children and adolescents. Implicit understanding was investigated by collecting their predictions of possible outcomes regarding the familiar causal system. Explicit understanding was investigated by collecting explanations for why those specific predictions were made. Degree of conviction in the prediction was

also collected, to allow the point of comparison with the adult literature. If children have an epistemic understanding of what different sources can be expected to know, then that understanding may impact on predictions, explanations, and their degree of conviction.

The relevance of factors that have been shown to potentially play a role in scientific reasoning, such as language ability, and gender, are examined. They are frequently not taken into account when investigating source reliability understanding in children (beyond including demographic information regarding the participant sample), even though they are known to impact reasoning in other related areas (such as scientific reasoning when it is assessed in the school environment, see section 1.1.3). Thus a third aim of the thesis was to investigate if and how source reliability was related to language ability and gender in relation to understanding of a familiar causal system.

## 1.4 EXPERIMENTAL PARADIGM

The causal system chosen for investigation needed to fulfil a number of requirements. It needed to be a familiar causal system, that even young participants understood. If the younger participants did not have a basic understanding of how the causal system worked, then they would not be able to make appropriate predictions or explanations regarding the system. It would therefore be difficult to assess the impact of source reliability, where observation of systematic changes in prediction and explanation dependent on source of information will provide evidence of epistemic awareness of source reliability. Furthermore, there needed to be information relating to how the causal system worked, that was not common knowledge among either children or adolescents.

The causal system that was chosen was motion on an incline, specifically a car travelling down an inclined track. Motion on an incline was one of the tasks chosen by Inhelder and Piaget (1958) to examine the growth of logical thinking from childhood to adolescence, so it fits the criteria that it is suitable for all age groups. It is also a topic that is learned in both primary and secondary school, to various degrees (Department for Education, 2015). Furthermore, although the impact of altering variables such as height of the incline, starting point on the incline, or surface friction on the incline, and distance travelled are well-known, the effect of

weight on distance travelled is not. That is, many children (and adults) have misconceptions regarding the effect of weight on distance travelled, and frequently think that weight *does* impact on distance travelled, even though it does not (Hast & Howe, 2012). This is the case even for secondary school children who learn about Newtonian mechanics at school, yet still report misconceptions regarding the impact of weight on motion (Mildenhall & Williams, 2001). The majority of children are likely to know that height, starting point on the incline, and surface friction affect how far the car travels, and are also likely to think that weight affects how far the car travels (Ferretti, Butterfield, Cahn, & Kerkman, 1985; Howe, Tolmie, & Rodgers, 1992; Hast & Howe, 2012; Hast & Howe, 2013).

These misconceptions made it possible to give participants 'unknown' information regarding the system at hand – that weight does not affect how far the car travels – from differentially reliable sources, and assess whether the source information was related to their reasoning regarding the causal system. The paradigm made it possible to ask participants for predictions, explanations, and degrees of conviction regarding how far they think the car will travel, for the different variables (weight, height, starting point, friction). Participants were only asked questions regarding variables in a single dimension, as children are more likely to make errors of understanding when variables interact (Ferretti et al., 1985). This is because a basic understanding of how the causal system works is desirable, so that systematic changes in understanding could in principle be observed.

Degree of conviction was measured using a rating scale that was similar to the scale used in Schlottmann and Anderson (1990) to assess children's understanding of expected value.

Prior to testing, participants were able to 'play' with the causal system, by allowing the car to run down the incline and observing how far it travelled for a number of trials. Information on how often children spontaneously and correctly assess the effect of specific variables, along with how frequently they assess scope of the system (least and furthest the car could travel), and repeat trials, was collected. Both children and adolescents frequently do not spontaneously engage control of variable testing (Cook et al., 2011) when left to 'play' with a system. This enables the study to conduct exploratory analyses examining what children might seek to

discover when faced with a conceptually familiar causal system, but specifically unfamiliar.

The study was designed to ask participants to make predictions, explanations, and report a rating of degree of conviction, for each variable of the causal system - car on an incline - where variables were height of the incline, starting point on the incline, surface friction, and weight, before and after they received information from differentially reliable sources, or no information. Participants were also asked to intervene on the system, where they observed that weight did not, in fact, affect how far the car travelled. After this they were asked to make a third set of predictions, explanations and report a rating of degree of conviction.

## 1.5 HYPOTHESES

### 1.5.1 IMPLICIT UNDERSTANDING OF THE CAUSAL SYSTEM (PREDICTION)

If source reliability is associated with participants' understanding of the causal system, then one would predict that participants will differentially utilise information from high and low sources when making predictions regarding the effect of weight on distance travelled down an incline. Participants who received information from high reliability sources would be more likely to make predictions that suggested they did not think weight affected how far the car travels compared with participants who hear no information. However, participants who received information from low reliability sources will be less likely to do this, and show no difference when compared which participants who received no information.

#### 1.5.1.1 AGE

If young children are showing epistemic awareness of what sources know, then one would predict that source reliability is related to predicting that weight does not have any relation to the distance the car travels, even in the younger participants. However, if epistemic awareness is slower to develop, particularly in more naturalistic environments, then the prediction is that an age difference will be identified, with older participants but not younger participants showing a source reliability effect.

Alternatively, given the inconclusive evidence, it could be predicted that source reliability is more relevant for younger participants than older participants, whose stronger prior beliefs regarding the familiar causal system will be more relevant.

After participants had the opportunity to intervene on the system, and witness that weight really did not affect how far the car travelled, if participants understand the implications of what they have observed, then the prediction is that source reliability will not be relevant to predictions about changing the weight of the car and the distance travelled. If participants did not understand the implications of the information, or did not fully understand how the system worked, then it is possible that some participants would continue to predict that weight does have an effect on distance travelled. This is more likely to occur with younger children, where their executive function skills are less well developed such that they cannot fully integrate the new information with their understanding of the system. They may also struggle to inhibit their strong prior beliefs regarding the effect of weight, even in the face of recent observational evidence to the contrary of their belief.

#### 1.5.1.2 Language

It is unclear from the existing literature whether language ability will be associated with making correct predictions. Given that generating predictions is part of scientific reasoning, and language ability has been shown to impact on scientific reasoning skills, the expectation is that better language development will be associated with generating more correct predictions regarding the effect of weight (an aspect of scientific reasoning).

#### 1.5.1.3 Degree of conviction

Regarding degree of conviction, based on research that has been conducted with adults, if there is an effect of source reliability, then the prediction is that participants who receive unexpected information from a high reliable source will be more convinced by their prediction, compared with participants who received no information. This would be in contrast to participants who received unexpected information from a low reliable source. The prediction is that participants receiving information from a low reliability source will not differ from participants who received no information.

### 1.5.2 Explicit understanding of the causal system (Explanation)

As with implicit understanding of the system as it related to predictions about the distance the car will travel, participants who receive information from the high reliability source will be more likely to incorporate that information into their explanations regarding the effect of weight on how far the car travels, compared with participants who received no information. It is predicted that this will be less likely for participants who received information from a low reliable source, in comparison with participants who received no information.

After participants observed that the information was 'true', essentially receiving evidence that was much more salient regarding the effect of weight, then it is predicted that there will be no differences in correct explanations related to source reliability. Instead, the differences will be between participants who had heard the true information (either from a high or a low reliable source) and those that did not. That is, being provided with a verbal explanation that explains what they have observed allows the generation of a better explanation for the event, not likely for participants who have heard no explanation.

#### 1.5.2.1 Age

The ability to generate appropriate explanations improves with age, particularly as language ability improves with age. Given this, it is expected that older participants will be more likely to provide correct explanations compared with younger participants. However, the provision of an appropriate explanation is more cognitively demanding than providing an appropriate prediction. This is because providing an explanation requires an explicit understanding of the mechanics of the causal system. As such, it is predicted that the ability to provide appropriate explanations will be more likely for older children.

#### 1.5.2.2 Language

It is more likely that generating explanations will be affected by language ability (compared with prediction). This is because generating appropriate explanations relies on fundamental language skills that generating predictions does not. As such, it is predicted that level of language skill will affect participants' ability to generate appropriate explanations.

#### 1.5.2.3 Degree of conviction

The previous literature suggests that providing an explanation reinforces both children and adults' beliefs regarding the causal system. As participants' explanations are guided by their predictions, it is likely that the degree of conviction will also show a relationship with their explanations. The prediction is that participants who receive information from high reliable sources will show a higher degree of conviction in their prediction, *and* will provide better explanations.

### 1.5.3 Gender

Previous research suggests that gender is related to scientific reasoning in that males are likely to perform more highly in science at school. Research also indicates that females tend to show better language ability. It has also been suggested that language ability is related to performance in science at school, although the gender differences in language, and in science performance are frequently quite small, particularly with younger children. As such, gender may influence both prediction and explanation. Specifically, if there are gender differences, then it would be expected that males will be more likely to provide correct predictions but females, with better language ability, could be predicted to provide better explanations to justify their decisions.

### 1.5.4 Understanding the causal system

In order for participants to demonstrate a preference for more reliable sources they need to be able to do a number of things. They need to decide to:

- use the information from reliable sources;
- understand the causal system,
- understand the implications of the information they receive regarding the causal system, *and*
- generate predictions/explanations regarding the causal system, based on this new information.

Therefore, it is important to assess participants' general understanding of the causal system. It is predicted that all age groups will be able to generate appropriate predictions regarding distance travelled when varying each of the variables that can be manipulated in the causal system.

## 1.6 THE FOLLOWING CHAPTERS

In order to address the goals of the thesis, testing was done in two primary and two secondary schools. Chapter 2 describes the general methodology used. Chapter 3 presents the results, and chapter 4 provides a general discussion of the findings in the study. The reference list, and appendix follows.

# 2 General Methodology

## 2.1 School information

### 2.1.1 Primary school recruitment

Before beginning recruitment, ethical consent to do research with minors (under 16 years) was gained from the Ethics Committee in the Department of Psychological Sciences, Birkbeck, University of London.

The primary schools came from two different boroughs. The choice of the first borough was opportunistic as it was near to the researcher's home. School A was contained within this borough. The second borough was chosen as it contained different socioeconomic status characteristics to the first. School B was contained within this borough.

A list of potential primary schools in the borough of interest was available from each borough's education department, for parents choosing primary schools for their children. School reception staff were then contacted and asked who would be the best person to contact regarding this matter. If possible, the head teachers were contacted directly as they were the people who would make the final decisions regarding whether a study could be run at their school. However, usually the schools requested that the information was sent to a generic school email, which would then be forwarded to the head teacher. Schools (N=32) were then contacted with a speculative email (see Appendix A) that was either addressed to the head teacher or generically addressed, depending on what information was given by the school reception team. An information sheet was included containing brief background information on the study as well as what would be required of the school and the participating pupils. This included all the researcher's contact details, as well as those of the study supervisor (see Appendix A). A phone call to reception was made following up on the email. This process resulted in two primary schools agreeing to participate, school A and school B.

#### 2.1.1.1 School A

Primary school A is a single form entry church school that serves a reasonably affluent area of London. Priority admission is given to parents who are regular worshippers. In comparison with the national average, differences of note include a

much higher proportion of pupils with English as a second language (41%), and fewer pupils with free school meals (see Table 2-1; Department for Education, 2015 - 16 Cohort). It is likely that the higher proportion of pupils with English as a second language is related to being within London, which is much more ethnically diverse than much of the UK. Free school meals are frequently used as a proxy for indicating low family income (Hobbs & Vignoles, 2010), and the small proportion of eligible pupils suggests that fewer pupils at school A come from deprived family environments. In the UK, students' educational progress in reading, writing and mathematics, is evaluated at age 6-7 (year 2), 10-11 (year 6), and 13-14 (year 9) using Standard Assessment Tests (SATS; National Curriculum Assessments, 2017). Records indicated that pupils in school A performed either well above average or average in the SATS taken in year 6, with the majority reaching the expected standard in English and mathematics (93%). The school A website provided curriculum-based information for parents, including a detailed breakdown of what students learn in all subjects in each year of primary school, as well as curriculum leaflets which briefly describe what the pupil will be learning in each term.

Table 2-1 Primary school A demographics and performance, compared with the national average.

| | **Primary School A** | **National Average (England)** |
|---|---|---|
| Yearly pupil intake | 30 | N/A |
| Admissions Criteria | Priority church applicants | N/A |
| Pupil/Teacher ratio | 19.6 | 20.5 |
| **Demographics** | | |
| Gender | 49% female | 49% female |
| English as a 2nd language | 41% | 20% |
| Special Educational Needs | 0.5% | 2.6% |
| Free School Meals | 4.0% | 25% |
| Absences | 2.2% | 4.0% |
| **Performance in Year 6 SATS[1]** | | |
| Reading | Well Above Average | N/A |
| Writing | Average | N/A |
| Maths | Average | N/A |
| % reaching expected standard in English/Maths | 93% | 61% |

[1]Standard Assessment Tests

#### SCHOOL A PARTICIPANT RECRUITMENT

When recruiting the participants in school A, the researcher provided the head teacher with copies of the information sheet for parents (Appendix A) and the parental consent form (Appendix B) to give to all the parents in each year group. The head teacher actively promoted the study to the parents in a newsletter, and following that, the information sheet and consent forms were handed to the students to take home to their parents. The school collected the signed consent forms and collated a list of students whose parents had consented. This list was given to the researcher along with the consent forms. Approximately two thirds of the students participated from each class. Details of the exact proportion of students who participated in the study per age group is recorded in the 'Participants' section, in Table 2-5. The consent of the student was solicited verbally at the beginning of each testing session. The school provided a small room to carry out the experiment, which overlooked the playground, which was noisy when testing coincided with students being on a break. Year groups were tested when they were available (determined by the year group teacher).

### 2.1.1.2 SCHOOL B

Primary school B is also a single form entry church school, although the area of London it serves is less affluent. Unlike school A, only 60% of students receive priority admission, when their parents are regular worshippers. The remainder of places are given to students based on distance from the school. Similar to school A, School B has a higher proportion of students with English as a second language (36%) than the national average (20%). However, unlike school A, they have many more students receiving free school meals (21% - similar to the national average), suggesting that a greater number of students in school B come from a more deprived family environment, compared with school A. Students in school B performed either well above average or average in reading, writing and mathematics, with most students in the school reaching the expected standard in English and mathematics SATS (74% - above the national average of 61%; Department for Education, 2015-16 cohort). See Table 2-2 for specific details. School B provides parents with details of their learning and teaching ethos, whereby they have a creative curriculum, seeking to use art and music within their everyday teaching. They also provide

detailed curricula for maths and English, as well as curriculum leaflets for each term which briefly describe what the student will be learning.

Table 2-2 Primary school B demographics and performance, compared with the national average.

| | **Primary School B** | **National Average (England)** |
|---|---|---|
| Yearly student intake | 30 | N/A |
| Admissions Criteria | 60% Priority Church applicants<br>40% Distance | N/A |
| Student/Teacher ratio | 20.3 | 20.5 |
| **Demographics** | | |
| Gender | 52% female | 49% female |
| English as a 2nd language | 36% | 20% |
| Special Educational Needs | 0% | 2.6% |
| Free School Meals | 21% | 25% |
| Absences | 2.7% | 4.0% |
| **Performance in Year 6 SATS[1]** | | |
| Reading | Well Above Average | N/A |
| Writing | Average | N/A |
| Maths | Well Above Average | N/A |
| % reaching expected standard in English/Maths | 74% | 61% |

[1]Standard Assessment Tests

### SCHOOL B PARTICIPANT RECRUITMENT

The researcher used a similar procedure for recruitment compared with school A, whereby she provided the head teacher with copies of the information sheet and parental consent form (see Appendix A and B). The class teachers handed them out to all the parents in each age group and collected the signed consent forms. These were returned to the researcher. The study was not promoted to the parents to the same extent compared with school A and a smaller proportion of students in school B signed up to participate in the study (see Table 2-5). The consent of the student was solicited verbally at the beginning of each testing session. The research took place both in a small office (where a teacher sometimes quietly worked during testing) or in an empty classroom. The classroom was occasionally noisy due to classroom activities in adjacent classrooms. Year groups were included when they were available (dictated by the head teacher).

### 2.1.2 SECONDARY SCHOOL RECRUITMENT

Before beginning secondary school recruitment, a second ethical approval to do research with minors (under 16 years) was gained from the Ethics Committee in the Department of Psychological Sciences, Birkbeck, University of London. The second ethics application was made because the age range of the participants had changed, as well as the addition of a reward for participation.

Both secondary schools were recruited via direct contact with a teacher within the science department. The researcher knew a student in school C (a girls' school), which led to direct contact with the head of psychology. This teacher was interested in the study, and solicited permission for the study to take place in the school from the head teacher, and agreed to be the main point of contact. A similar process took place with the recruitment of School D (a boys' school), where the researcher's supervisor knew a chemistry teacher who also solicited the head teacher for permission for the study to take place at the school, as well as agreeing to be the main point of contact.

#### 2.1.2.1 SCHOOL C

Secondary school C is a seven form entry girls' Academy, serving both affluent and non-affluent adjacent areas of London. It also has a coeducational sixth form. It serves a large ethnic population where around 40-50% of inhabitants were born outside England according to the 2011 census. Admission is mainly based on distance between home and school, although 25% of the yearly student intake is based on performance on the Year 6 SATS for students living in the appropriate borough. It is a science specialist school, providing an enriched science and mathematics curriculum. Given the large minority ethnic population from which they draw, unsurprisingly a large proportion of students have English as a second language, which is much greater than the national average (60% versus 16%). There is also a greater proportion than the national average in England of students receiving free school meals (45% versus 29%). School C is actively engaged in trying to reduce the attainment gap between students from disadvantaged backgrounds and report on number of students receiving pupil premium (government funding for disadvantaged students, based on receipt of free school meals). They indicate that they have successfully managed to reduce the gap between students who received the pupil premium and their peers from 27% in 2014 to 18% in 2016

(information from school website, accessed on 29th March, 2017). Students in school C performed above average at GCSE, with 70% getting C or higher in English/maths (59% national average). 54% of students receiving the pupil premium received five A* to C GCSE grades and their non-disadvantaged peers received 72% A* to C GCSE grades (information from school website, accessed on 29th March, 2017). However, the average results for A-level were slightly below average (C- versus C+). Information accessed on Find and Compare Schools in England (Department for Education, 2015-16 cohort; See Table 2-3 for details).

Table 2-3 Secondary school C demographics and performance, compared with the national average.

| | **Secondary School C** | **National Average (England)** |
|---|---|---|
| Yearly student intake | 280 | N/A |
| Admissions Criteria | 25% Performance on Year 6 SATS[1] within borough<br>75% Distance from home | N/A |
| Student/Teacher ratio | 15.2 | 15.3 |
| **Demographics** | | |
| Gender | 99% female (boys in sixth form) | 49% female |
| English as a 2nd language | 60% | 16% |
| Special Educational Needs | 1.20% | 3.90% |
| Free School Meals | 45% | 29% |
| Absences | 5.7% | 5.3% |
| **Performance** | | |
| Overall Performance at GCSE | Above average | N/A |
| GCSE C or higher in English/Maths | 70% | 59% |
| Overall Performance at A-Level | Average | N/A |
| Average Result | C- | C+ |
| % with AAB or better | 6.80% | 17% |

[1]Standard Assessment Tests

### SCHOOL C PARTICIPANT RECRUITMENT

When recruiting the participants in school C, the researcher provided the contact teacher with a pack for students aged 12-15, containing the information sheet for parents (Appendix A), the parental consent form (Appendix B), and a consent form for the student (Appendix C) as the majority of students in this study would be over 13 (where signed consent is required). The written consent of the few students who

were under 13 was collected for the sake of consistency. Students aged 16 – 17 did not need parental consent to participate in the study, so they were provided with their own information sheet and consent form (Appendix C).

In order to encourage participation, secondary school students were also offered a £5 gift voucher for participating, which was made known to them as part of the recruitment process by the contact teacher, as well as being mentioned in the information sheet. The contact teacher promoted the study both in classrooms, and in school assemblies (which covered the entire year group). Any student who was interested in participating would then seek out this teacher. She signed up the 12- to 15-year-old participants for testing as soon as they returned the signed parental consent forms, as well as providing their own signed consent form. Participants over 16 needed only to provide a consent form to be signed up for the research. The teacher used an appointment schedule provided by the researcher. The teacher also reminded participants when they were due to take part. Without this help, far fewer participants would have been involved (in fact, the contact teacher became very busy with school work when Year 8 were taking part, which is noticeable in the fewer number of participants in that year group). Verbal consent was also solicited from the participants at the beginning of each session.

The school only permitted the research to take place in break times, lunch, and after school. However, students were allowed to leave class on occasion, which was negotiated by the teacher contact. An unused science lab was provided for the majority of the research, shifting to an unused classroom for the remaining participants. These rooms were not very noisy, even when students were on a break.

#### 2.1.2.2 SCHOOL D

Secondary School D is a four form entry boys' Academy, serving less affluent areas of London. Admission is mainly based on distance, with 10% musical aptitude, as this school has a long-standing musical tradition. It also has a mixed gender sixth form. Around 30% of inhabitants of the local area were born outside of England according to the 2011 census which is reflected in the number of students with English as a second language (49%). Even more students than secondary school C receive free school meals (57%), which is almost double the national average. Students in school D performed below average overall in their GCSEs, but above

average when just considering English and maths, where 67% scored C or higher. School D also reports on the impact of receiving the pupil premium, although provides slightly different data than school C. They report a gap of 28% for disadvantaged compared with non-disadvantaged students, achieving at least a grade C in GCSE English/maths (pupil premium students - 59% versus peers - 87%). This is in line with the national average - 27%. School D also performs lower than average at A-level (D+ versus C+). Information accessed on Find and Compare Schools in England (2015-16 cohort; See Table 2-4 for details).

Table 2-4 Secondary school D demographics and performance, compared with the national average.

| | **School D** | **National Average (England)** |
|---|---|---|
| Yearly pupil intake | 130 | N/A |
| Admissions Criteria | 10% Musical Aptitude<br>90% Distance from home | N/A |
| Pupil/Teacher ratio | 13:1 | 15.3 |
| **Demographics** | | |
| Gender | 88% male (girls in sixth form) | 49% female |
| English as a 2nd language | 49% | 16% |
| Special Educational Needs | 2.10% | 3.90% |
| Free School Meals | 57% | 29% |
| Absences | 5% | 5.3% |
| **Performance** | | |
| Overall Performance at GCSE | Below average | N/A |
| GCSE C or higher in English/Maths | 67% | 59% |
| Overall Performance at A-Level | Below Average | N/A |
| Average Result | D+ | C+ |
| % with AAB or better | 2.70% | 17% |

### SCHOOL D PARTICIPANT RECRUITMENT

The procedure for recruitment in school D differed slightly from school C. The contact teacher in school C arranged permission from the head teacher, and handed out packs with information sheets and consent forms for both parents and pupils (see Appendices A to C) to form teachers to hand out in class. This differed from school C, where only pupils who volunteered were given the information sheets and consent forms. Pupils who wanted to participate would return their signed consent forms (theirs and their parents) to the science office, and sign up for an appointment

from a schedule that was left near the contact teacher's desk in the science office. As there was no single room available for the duration of the research, the contact teacher would book rooms, and a sign would be left on the science office door saying in what room the research would take place on a particular day. Verbal consent was also solicited from the participants at the beginning of each session.

Recruitment in school D was quite challenging. All the teachers were busy, including the contact teacher, and did not have much time to help with recruitment, and scheduling of participants. If the participants were not reminded when to attend a research session, they often would not show up, and locating the missing student was difficult. The researcher frequently had to collect the participant, if she could find them, to ensure participation. As a result of that there were many fewer participants in school D, where many more potential participants signed up than participated.

The school only permitted the research to take place in break times, lunch, and after school. Unused science labs were used for testing, which were sometimes noisy at break time or after school when many pupils were departing school at the same time.

## 2.2 PARTICIPANTS

The research involved participants ranging in age from six to 17 years. The mean age (S.D.), age range, gender as a proportion of females in the age group, number of participants and percentage of the school year group that participated are recorded in Table 2-5. The following participants had some data collected but were not included in the study. One child in school A (10-11 years) failed to provide appropriate responses, and failed to complete testing. One child in school C (16-17 years), and one in school D (12-13 years) provided incomplete data (doing only the first testing session).

Table 2-5 Participant characteristics: mean and S.D. by age group; age range; number participating per age group; number of participating females; the proportion of the year group participating.

| Age group | School | Mean | S.D. | Range | Total N | (N female) | % of Year |
|---|---|---|---|---|---|---|---|
| 6-7 years | A | 7.2 | 0.31 | 6.6 – 7.5 | 20 | 10 | 67 |
| | B | 7.3 | 0.40 | 6.8 – 7.8 | 13 | 8 | 43 |
| | Overall | 7.2 | 0.35 | 6.6 – 7.8 | 33 | 18 | 55 |
| 8-9 years | A | 9.1 | 0.36 | 8.5 – 9.5 | 20 | 10 | 67 |
| | B | 9.2 | 0.19 | 8.9 – 9.4 | 6 | 3 | 20 |
| | Overall | 9.1 | 0.33 | 8.5 – 9.5 | 26 | 13 | 44 |
| 10-11 years | A | 11.3 | 0.28 | 10.9 – 11.8 | 18 | 8 | 60 |
| | B | 11.4 | 0.42 | 10.9 – 11.8 | 9 | 4 | 30 |
| | Overall | 11.3 | 0.33 | 10.9 -11.8 | 27 | 12 | 45 |
| 12-13 years | C | 13.3 | 0.23 | 13.0 – 13.7 | 11 | 11 | 4 |
| | D | 12.9 | 0.27 | 12.6 – 13.3 | 9 | All male | 7 |
| | Overall | 13.1 | 0.31 | 12.6 - 13.7 | 20 | 11 | 6 |
| 14-15 years | C | 15.1 | 0.36 | 14.5 – 15.7 | 29 | 29 | 10 |
| | D | 15.2 | 0.38 | 14.6 – 15.4 | 4 | All male | 3 |
| | Overall | 15.1 | 0.36 | 14.5 - 15.7 | 33 | 29 | 7 |
| 16-17 years | C | 17.0 | 0.38 | 16.4 – 17.9 | 22 | 22 | 8 |
| | D | 17.0 | 0.35 | 16.5 – 17.4 | 8 | 6 | 6 |
| | Overall | 17.0 | 0.36 | 16.4 - 17.9 | 30 | 28 | 7 |

A further nine participants were removed from the analysis as they had receptive vocabulary scores (measured using British Picture Vocabulary Scale; see section 2.5.4) more than 2 S.D. below the mean (see Table 2-6).

Table 2-6 Age group, BPVS[1] score, gender, and school of the nine participants who were removed from the analysis.

| Age | BPVS[1] Score | Gender | School |
|---|---|---|---|
| 8-9 Years | 67 | Female | Primary School B |
| 10-11 Years | 62 | Female | Primary School B |
| 12-13 Years | 68 | Female | Secondary School C |
| 12-13 Years | 70 | Male | Secondary School D |
| 14-15 Years | 48 | Female | Secondary School C |
| 16-17 Years | 56 | Female | Secondary School C |
| 16-17 Years | 58 | Female | Secondary School C |
| 16-17 Years | 59 | Female | Secondary School C |
| 16-17 Years | 45 | Female | Secondary School D |

[1] British Picture Vocabulary Scale

## 2.3 APPARATUS

### CARS ON AN INCLINE GAME

The familiar causal system that was used addressed participants' understanding of motion on an inclined plane. This was created using cars that could change weight, and start at different heights, starting points, and with different surface frictions on the inclined planes. Most school children (and many adults) think that weight *does* affect how far the car travels (even though it does not), younger children thinking that lighter vehicles travel farther and older children that heavier vehicles travel farther (Hast & Howe, 2012). These misconceptions allow children to receive unexpected 'expert' information from more, or less, reliable sources.

Furthermore, the majority of primary school pupils know that height, starting point on the incline, and surface friction affect how far a car travels (Ferretti et al., 1985; Howe et al., 1992; Hast & Howe, 2012; Hast & Howe, 2013), and asking about it allows for assessment of understanding of the causal system.

The 'Car on an Incline' game assesses understanding of the causal system associated with how far a car could travel down an incline, changing friction on the incline, height of the incline, starting point on the incline, and weight of the car. The causal variables affecting distance travelled are height of the incline, starting point on the

incline, and surface friction of the incline. The weight of the car is not a causal variable as it does not affect how far the car travels.

The game consisted of a frame upon which four inclined ramps rested in a row. Each incline measured 78cm x 10cm, with 0.5cm raised sides and was made of wood with a flexible plastic section at the lower end (this allowed the incline to smoothly segue onto the floor). Each incline rested on an adjustable bar connected to the frame (50cm wide) that could be raised/lowered to three equally spaced positions (high - 20.5cm, medium - 15.5cm, and low - 10.5cm, from the floor). Three of the inclines had observably different surface friction in varying degrees; one smooth (shiny sticky backed plastic), one medium (slightly textured wallpaper), and one a rough (very textured wallpaper). There was an extra incline identical to the medium friction surface ramp, included to remind participants where the car went on the standard setup (see below).

Each incline had three starting points (high - 56cm, medium - 43.5cm, and low - 31cm, from the bottom), where a gate could be inserted at a starting point, and removed to allow the car to travel down the incline onto the track.

A 13.5cm Burago BMW Cabriolet model car was used with three small equally weighted bags filled with lead shot, of around 2.5cm in diameter and small enough to fit in the back seat of the (open-topped) car. The light car contained one bag, the medium-weight car had two, and the heavy car had three bags.

The track was drawn onto cream coloured canvas, 175cm long by 70cm wide and equally divided into seven boxes, numbered 1-7 on the left-hand side, each 25cm long by 70cm wide. When the game was in use, the track was adhered to the ground with sticky tape at the four corners and along the long sides to make sure it remained flat at all times.

The game was calibrated so that the medium weight car arrived in the number 4 box when it started on the medium surface (friction), medium starting point (on the incline), medium height (of incline). This was referred to as the standard setup. The fourth incline had a medium friction surface, and was set at the medium height, with a gate at the medium starting point. A second car, identical to the first, with two bags inside, was left at the side of the track beside number 4 to remind participants where

the car would land when the set up was in the standard position. A diagrammatic representation of the standard setup can be seen in Figure 2-1.

When height or starting point were on the low position, and everything else was on medium position, the car landed in number 3. When height or starting point were on the high position, and everything else in medium position, the car landed on number 5. However, for surface friction, although the car landed in number 3 for the rough surface (similar to above), it landed in number 4 for both the medium and smooth surface positions (with the car landing slightly higher in the box for smoother surface position).

The car landed in number 4 for the light, medium and heavy weight car, demonstrating that weight does not affect how far the car travels. The light car did travel a little further in number 4, compared with the medium or heavy car, but this was difficult to notice unless many trials were completed.

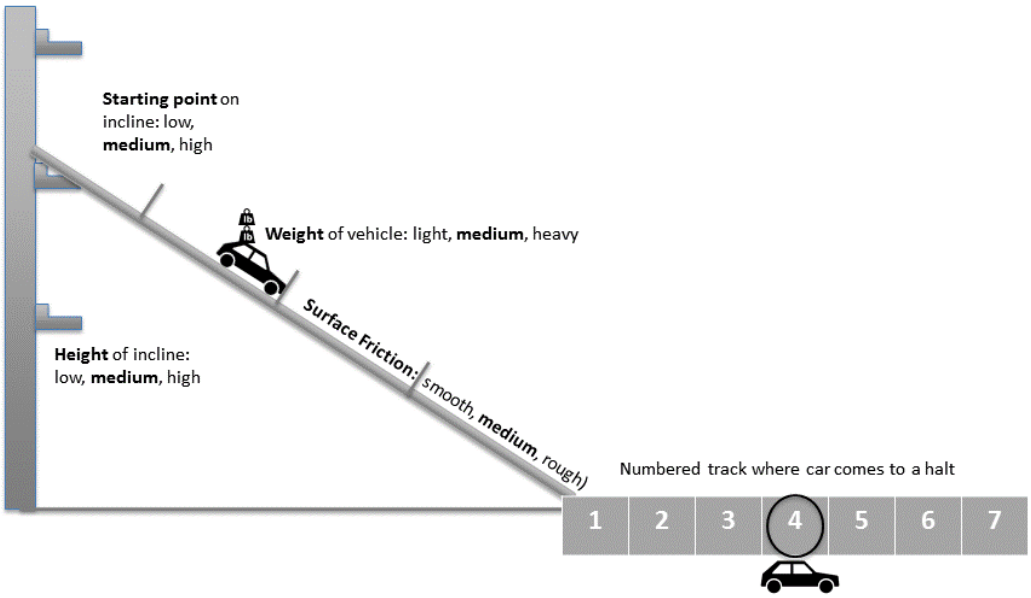


Figure 2-1 Car on an incline game in the standard set up.

### STICK SCALE

The stick scale used to collect ratings had a wooden base, 26cm (length) x 4cm (depth) x 5cm (height). Protruding from the base were seven differing lengths of

wooden dowel in height order. The shortest dowel was 2.5cm. They increased in height by 1cm, with the tallest dowel being 8.5cm. There were also visual reminders of what each end of the rating scale represented. The guessing end of the scale (the smallest dowel) was represented by a cartoon figure flipping a coin with a thought bubble containing images of a coin showing both heads and tails and two '??'. The completely sure end was represented by a cartoon figure with a thought bubble containing an image of the sun. See Figure 2-2 for a diagrammatic representation of the stick scale.

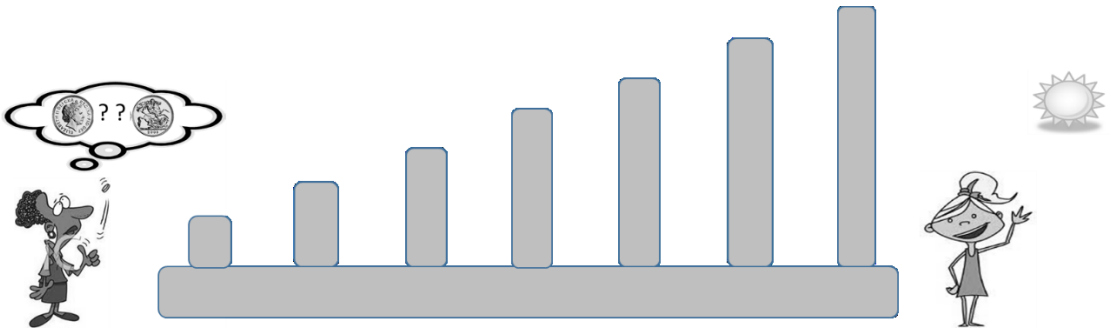


Figure 2-2 Diagrammatic representation of the stick scale with examples of the visual reminder to what each end of the scale represented.

The primary school participants' responses were video recorded using a Macbook, an Olympus LS-20 PCM Digital Recorder, or an IPad Air 2. For secondary school participants, only audio was recorded, using Voice Record Pro (an app for IPad).

## 2.4 DESIGN

## 2.5 INDEPENDENT VARIABLES

Source reliability, age and receptive vocabulary were between-subject factors, and type of variable and time of testing were within-subject factors.

### 2.5.1 SOURCE RELIABILITY

There were a high and a low reliability source condition, and a 'no information' condition. In the high and low source reliability conditions, participants were told by the researcher:

*"I was discussing this with a [science or physics teacher/nursery child] last week. They said that they thought that the weight of the car does not make a difference to how far the car travels. What do you think?"*

A *science teacher* was used as the high reliability source for primary school participants. Not only do children tend to trust what adults tell them (e.g. Harris & Koenig, 2006), but they are also likely to recognise expertise when evaluating whether to trust a statement (Kushnir, Vredenburg, & Schneider, 2013). Fitneva (2010) also found that six-year-olds generally discriminate between adult and child information, perceiving adults as more knowledgeable, but four-year-olds do not. Furthermore, Yeo and Tan (2010) found that receiving relevant information from an authoritative source, such as a teacher, is likely to improve students' learning in secondary school, suggesting that teachers can function as high reliability sources even with older children. Bråten et al. (2012) also found that secondary school students were likely to prefer justification by authority, over justification by multiple sources and personal justification, when asked about justification of knowledge in science. Children in the UK studying the national curriculum (Department for Education, 2014), also learn about forces and motion in their science class from around five years old, so they have had exposure to the relevant concepts regarding the behaviour of cars on inclines in their 'science' class.

However, while participants should regard a science teacher as being an expert in science, and therefore knowledgeable regarding the variables under consideration, it is likely the participants will not consider the science teacher completely reliable

(never incorrect). This is because children tend to defer to adults as being reliable sources of information regarding particular topics (e.g., Lutz & Kyle, 2002), *unless* they consider themselves 'experts' (e.g. regarding the action of particular toys; Vanderborght & Jaswal, 2009). It is possible that even young children would feel that they are 'knowledgeable' about the effects of the variables under consideration here, which may impact on the degree to which they think the science teacher is reliable.

A *physics teacher* was used as the reliable source for secondary school participants. This was because pilot work indicated that adults appear not to rate the science teacher as reliable (unlike primary school participants), and it was felt that this might be the case for secondary school participants as well. This was especially relevant as they do 'science' under more specific labels of physics, chemistry, and biology (see Department for Education, 2013). Given that the intention was for the participants to associate the reliable 'expert' source with the causal (physical) system directly, it was decided that the 'physics teacher' would be a better reliable source.

A nursery school child (three- to four-year-olds) was used as the low reliability source for both primary and secondary school participants. Children regard 'children' as being more unreliable than adults, as having less knowledge than adults (the younger the child, the less knowledgeable), and preferring to learn from adults (e.g. Taylor, Cartwright & Bowden, 1991). As such, participants should regard a child younger than them as knowing less, and therefore not a reliable source of information.

The statements used to manipulate source reliability were similar to Hahn et al. (2005), who found source reliability differences in adults. Younger participants have also been shown to be capable of discriminating between sources. Fitneva (2001) used brief statements indicating the reliability of the source via grammatical implication. As such, information regarding source reliability was given to the participants verbally.

A 'no information condition' was also included, where participants were given no information regarding the effect of weight on distance travelled. The 'no information' condition functioned as a control, where weight was made pertinent, but they were given no further information regarding its effects. The intention was

to avoid the problem, particularly with younger participants, that they might not pay attention to weight, because weight had not been made pertinent (as it has been in the other two conditions). At the point where the information regarding weight would be revealed to the child in the high and low reliability conditions, the child was asked:

*"So what effect do you think, say, weight will have on how far the car travels?"*

To ensure that the source reliability manipulation worked, all the participants were asked the following question, at the end of the study:

*"Out of 10, how often do you think a [nursery child; science/physics teacher] would be right if you ask them a question?"*

Children as young as four years of age are likely to be able to understand and identify who would know about adult specific and child specific knowledge (Fitneva, 2010). Primary school participants had no problems answering this question. Secondary school participants usually asked what kind of question, at which point the researcher told them the kind of question one might ask in a primary school science class.

### 2.5.2 WEIGHT AND CAUSAL VARIABLES

Four variables that *could* impact on the causal system were manipulated in the '*cars on an incline*' game (see Figure 2-1). They were *height* of the incline, *starting point* on the incline, *surface friction*, and *weight* of the car. The first three are causal variables - changing them impacts on how far the car travels. However, weight is not a causal variable - changing the weight of the car makes little or no difference to how far the car travels. Height, starting point, surface friction and weight could be varied, each having three possible setups as follows: height (high, medium, low); starting point (high, medium, low); surface friction (smooth, medium, rough); and weight (heavy, medium, light). For the causal variables, the high or smooth setup equated to the furthest position the car could travel from that position, the medium setup, equating to the in-between distance from that position, and low or rough setups, equating to the least distance the car would travel in that position. For weight, the car would travel to approximately the same place, regardless of how heavy the car was.

### 2.5.3 TIME POINT

Data were collected from the participants at three time points. At baseline - just after participants had done their practice trials; post reliability information – just after participants had received information (or not), regarding the effect of weight, from differentially reliable sources; and post intervention – just after participants have intervened on the system and witnessed that weight does not have an effect on how far the car travels.

### 2.5.4 RECEPTIVE VOCABULARY

Participants were given a receptive vocabulary assessment – the British Picture Vocabulary Scale (BPVS; Dunn & Dunn, 2009).

The Peabody Picture Vocabulary Test (PPVT, Dunn & Dunn, 1981) is a widely used measure of receptive vocabulary. Validity studies in the United States have shown that it correlates positively with other vocabulary tests and with individual intelligence tests (see Robertson & Eisenberg, 1981, for a review). The British Picture Vocabulary Scale (BPVS, Dunn & Dunn, Whetton & Pintillie, 1997) was based on the PPVT, with standardisation on a British national sample, and the drawings reworked for a better gender and ethnic balance as well as to remove content that was not representative of British culture.

The BPVS was particularly useful as it is appropriate for the entire age range included in this study (6-17 years), and it can be implemented fairly quickly (15-20 minutes). Furthermore, vocabulary is a good index of school success (Dale & Reichert, 1957) and is a big contributor in measures of intelligence (Elliot, 1982).

The participants were told a word and asked which of four numbered pictures matched the word. They had four practice trials, where they were told if they were correct or not. For subsequent words they were not told if they were correct. The words to identify were easy to start with, and got progressively more difficult. There were 32 potential picture sets. The task ended when participants failed to identify the picture that correctly matched the word four out of six times. The final correctly identified word picture pair is known as the ceiling item. The errors made by the participant were recorded, along with the ceiling item.

The raw scores were calculated by subtracting the number of errors from the ceiling item. These raw scores were converted to standardised scores, by age by using the tables provided in the manual for the BPVS (Dunn, & Dunn, 1981).

### 2.5.5 Age

In primary school, participants were selected from three different year groups, Year 2 (six- to seven-year-olds), Year 4 (eight to nine-year-olds), and Year 6 (10 to 11-year-olds ). In secondary school, participants were selected from three different year groups, Year 8 (12- to 13-year-olds), Year 10 (14- to 15-year-olds), and Year 12 (16- to 17-year-olds).

## 2.6 Dependent variables

### 2.6.1 Practice trial measures

Participants' choices of practice trials were recorded. This was done to examine what sort of knowledge the participants sought to gain when doing the practice trials. Whether or not the participant did the following types of trials was counted.

#### Sequential fair test trials

A fair test is when only one variable was changed at a time. For example, a fair test on height would be done where the high and low set up for height was compared, and the other three variables remained constant. A fair test was counted as such when the two set ups were compared sequentially. As there were six trials, it was possible for a participant to do a non-consecutive fair test. However, as sequential fair test set ups suggested the greatest likelihood of the participant attempting to compare distance travelled for different set ups, only they were counted.

#### Extreme set up trials

The extreme set ups of the causal system are the set ups where the car would travel the least and furthest on the track. Whether the participant did an extreme set up trial at either or both ends were counted. The car would land in Box. 6 for high height, high starting point, and smooth surface friction (with any weight), and in Box. 2 for low height, low starting point, and rough surface friction. Finding these points would give the participant more information regarding the causal system as it dictates the boundary of possible distance travelled by the car.

#### NUMBER OF REPEAT TRIALS

Participants could either repeat a trial because they may have forgotten what they had already done, or to confirm what they have observed. Number of repeat trials were counted.

### 2.6.2 VARIABLE PREDICTION MEASURE

The variable prediction measure allowed for assessment of whether participants correctly understood the effect of weight and the causal variables (starting point on slope, height of slope, surface friction).

The difference between participants' predictions for the high and low set up for each variable were calculated by subtracting the predicted distance from the low set up from the predicted distance for the high set up for that variable. The high set up was equivalent to the one where the car would travel furthest, and the low set up where the car would travel the least far.

#### WEIGHT PREDICTION

For weight, if this difference was 0, then the participant's understanding of the effect of weight was considered correct. This is because predicting that the heavy (high) and light (low) car will land in the same box (hence difference = 0) suggested the participant believed weight did not have an effect on distance travelled (in this instance). If the absolute difference was bigger than 0, then the prediction was incorrect. A positive difference suggested the participant thought the heavy car would travel further, and a negative difference suggested the participant thought the light car would travel further.

#### CAUSAL VARIABLES PREDICTION

For the causal variables, if the difference was positive, then the participant's understanding of the effect of the causal variable would be considered correct as they would be predicting that the high set up for that variable will travel further than the low set up. There are two types of causal variable prediction errors. The first was when the prediction difference was negative. An example of this type of error would be the participant predicting that the car starting at the lower starting point on the slope would go further on the track than if it started at the higher starting point. The second type of prediction error would be when participants have a prediction difference of zero. This suggested that the participant believed that a

causal variable had no effect (in this instance). It was also possible that the participant believed that the causal variable only has a *small* causal effect within that causal system, resulting in the participant predicting they both land in the same box. However, this did not occur in the cars on an incline game, and all causal variable prediction differences of zero were considered to be prediction errors.

Each participant made three sets of predictions for the high and low set up per weight/causal variable, at each time of testing.

### 2.6.3 Degree of conviction measure

The degree of conviction measure allowed for assessment of how sure participants were regarding their weight predictions. The scale ranged from 1-7, where 1 indicated that they reported that were just guessing, and 7 indicated that they were completely sure. Participants' degree of conviction rating for each prediction regarding the heavy and light set up for weight were combined to create a total degree of conviction rating for the weight prediction.

### 2.6.4 Weight explanation measure

During the data collection phase participants were required to state how far they thought the car would travel for a particular set up (prediction), how sure they were (degree of conviction), *and* why they thought that (explanation). They did this for both weight (non-causal variable) and the causal variables.

The audio recordings regarding explanations for the effect of the weight were transcribed using a professional transcriber. This was done at baseline, post reliability information, and post intervention. The explanation data were uploaded into the Atlas.ti software, designed to assist in analysing large bodies of textual information. This allows easily generated sets of explanations (e.g. age six to seven weight explanations), and all weight explanations were coded, with explanations scored depending on how well they mapped onto the correct generic explanation for weight.

The correct generic explanation for weight is as follows:

"Change in degree (heaviness) does not affect speed (stays the same) so it does not affect distance (stays the same)"

The participants scored according to the following criteria

0 - If they did not provide an explanation. For example - don't know; it just is; I can tell by looking; I just saw it (usually said after witnessing weight not having an effect).

1 - If their explanation provided was incorrect. For example - saying weight does have an effect; doesn't make sense (even if on right track); explanation involves other variables (e.g. saying friction/height/starting point play role when talking about weight).

2 - If their explanation was inadequate, missing information or very unclear. For example - saying "it is heavy"; missing out any other pertinent information; or off base in some way.

3 - If they provided a correct specific explanation. For example - saying that all degrees of weight (light, medium and heavy) will land in the same place. E.g. "they will all land in 4".

4 - If they provide a correct general explanation, where it is claimed that degree (of weight) does not affect how far the car travels (the latter may be implied). For example - saying "weight doesn't make a difference (to how far the car travels)". Weight does not make much of a difference was also accepted as there is a very small difference in reality.

5 - If they provided a correct explanation where degree, speed, and distance are all taken into account when explaining the prediction. Sometimes degree will be implied rather than stated explicitly. For example – saying it (weight) does not affect speed, so it does not affect how far the car travels.

Participants generated two explanations, one for the heavy and one for the light set up. As participants frequently treated these two explanations as two components of a single explanation both these explanations were treated as one for coding purposes. See Table 2-7 for examples of coded explanations for 0-3+ (there were very few explanations accorded higher scores than three).

Table 2-7 Sample quotes coded for ranking 0-3+ in the explanation scoring system.

| Explanation Score | Sample quote |
|---|---|
| 0 | -I'm not sure, I'm not too sure. I don't know<br><br>-I can see it going there<br><br>-I don't know I just guess. Just a random guess |
| 1 | -Because it's quite light and not very heavy it would go slow<br><br>-Because the car is heavier and… yes, the car is heavier<br><br>-Because it is heavier now, so it's less to go down |
| 2 | -Cause of the science teacher and about the weight and because it was on medium so I think it will go to 4<br><br>-Because it has one beanbag but it has two middle bits, so you add them both together they might make 4<br><br>-Because I would think that weight would make it go further, but obviously with the teacher and a bit more confused |
| 3+ | -Because of, well, it's like a science teacher said that weight doesn't really make much difference, so- and its medium high and medium surface, so I thought it would go to 4<br><br>-Because the weight doesn't matter. And if it has medium thing… a medium roughness, medium starting point and medium weight, it will go to 4, because the weight doesn't really matter<br><br>-Well, as the physics teacher said that weight does not affect the car's speed, then yes it would go to 4 |

## 2.7 PROCEDURE

All the trials took place in a private room within the participant's school. There were two sessions, 1-10 days apart. The researcher sat on a small chair behind the apparatus, and the participant sat facing the researcher, with a good view of the 'cars on inclines' game, and the numbered track. A procedure timeline is included below in Table 2-8.

Table 2-8 Main procedure timeline for the studies.

**Session 1 (10-20 minutes)**

- Welcome
  - Participant welcomed to the study, shown the apparatus, and asked if they wish to participate.
- Language Test
  - British Picture Vocabulary Scale.
- Introduction Phase
  - Participant's attention directed to variables that can change, and how they change, in the cars on an incline game.
  - Participant shown standard setup and witness that the car lands on box number 4 (repeated twice), with the standard setup.
- Practice Trials
  - Participant performs six practice trials (of their own choice)
- Training Phase
  - Participant told what questions they would be asked.
  - Participant told how to use the degree of conviction scale.
  - Participant practice using the degree of conviction scale.
- Test Phase I (Baseline)
  - Participant asked how far they thought the car would travel for the heaviest and lightest car, and the high and low positions for each of the causal variables.
  - Participant asked how sure they were, regarding each prediction.
  - Participant asked why they thought that, regarding each prediction.

A break of 1-10 days occurred between sessions.

**Session 2 (10-15 minutes)**

- Welcome
  - Participant welcomed to the study and asked if they wish to participate again.
  - Participant reminded of how game worked, and where car landed when on the standard setup.
- Reliability Information
  - Participant either told new information regarding the causal system that came from science/physics teacher or nursery child, or received no information.
- Test Phase II (Post Information)
  - Same as Test Phase I
- Intervention
  - Participant performed a fair test on weight, seeing how far the car travelled for the heavy medium and lightweight car. They did this twice.
- Test Phase III (Post Intervention)
  - Same as Test Phase I
  - Participant asked to give reliability ratings for science/physics teacher and nursery child.

## Session 1 (10-20 minutes)

### Language test

Firstly, the British Picture Vocabulary Scale (Dunn & Dunn, 1981) task was administered.

### Introduction phase

After the language test, participants were introduced to the 'cars on an incline' game. The researcher demonstrated how the game worked, touching the surfaces of the inclines and directing attention to the smooth, medium and rough surfaces, noting that there were high, medium, and low starting points, changed using gates. The inclines were raised up and down to demonstrate the low, medium, and high height positions. Finally, the researcher demonstrated how the weight of the car could be changed by putting different numbers of weights in the back seat. The car with three weights was referred to as the heavy weight car, two weights as the medium weight car, and one weight as the light weight car.

The researcher then pointed out that if the game was set up with the car on the medium surface, medium starting point, medium height, and medium weight then the car stopped in the middle of the track in the box number 4. The researcher then released the car so that the participant witnessed the car the landing where it was supposed to. This was repeated twice. The car was then left at the side of the number 4 box to remind the participant of where it stops in the standard set up.

### Practice trials

Following the introduction to the apparatus, the participant was allowed to play with the game, with no intervention from the researcher. The participant was told they could do anything they want, and that they have six practice trials.

### Training phase

In the training phase, the researcher returned the game to the standard set up. Then told the participant that the setup would be changed around, and that they would to be asked three questions, which were:

I. Where they think the car will land, where they would report the number of the box they think it would land in. The participant was then reminded that if the game was set up with medium weight car/height/starting point/friction (standard set up) then the car would land in box number 4.
II. How sure are they are about their prediction regarding how far the car would travel. They are also shown the stick scale at this point.
III. Why they think that.

The participant was then introduced to the stick scale. Their attention was drawn to the increasing height of sticks. They were told that the higher sticks mean they are more sure, and the lower sticks mean they are less sure. Then they were told the highest stick means they are completely sure.

To test their understanding, the researcher asked them how sure they were that the sun is going to come up tomorrow and that there is going to be another day. Their attention was drawn to the fact that this has happened during the entire lives of family members, for millions of years. The participant was judged to understand when they agreed that they would be very, very, sure. The researcher then referred to the lowest stick, and told the participant it means they are completely guessing. The researcher and participant discuss tossing a coin, and the fact that they would have to guess if it was head or tails. The participant was judged to understand when they agreed that they would be guessing. To remind the participant of which end is which, a cartoon picture of a person imagining a sun and a flipping coin was attached to the appropriate end of the stick scale.

Participants were then asked a series of questions to check they knew how to use the scale appropriately. The questions were designed to get them to use the middle of the scale as pilot studies suggest younger participants will often ignore the middle if they are not prompted. Primary school participants were usually asked questions such as 'how sure are you that this rubber will bounce' or 'how sure are you that if I push this toy car it will drive off the table' to get them to use the middle of the scale. Secondary school participants were asked questions regarding the weather such as 'how sure are you that it will rain tomorrow'. This questioning continued until participants used the middle of the scale to indicate their degree of conviction.

### Test phase I

The researcher, starting with a standard set up, asked the participant about each of the variables in turn, with weight always going first, and height/starting point/friction in a different order at each time of testing. For each variable, the higher and lower setup was demonstrated, where the other three variables were kept constant (in the medium position) and the participant was asked for their prediction regarding the distance the car would travel, how sure they were, and why did they think that (described above). The order of questioning regarding the higher and lower set up was systematically varied.

## Session 2 (10-15 minutes)

### Reliability information

On arrival, the participant was told they were going to do the same thing as last time, and was reminded of how the stick scale worked, and that they would get another chance to play the game later. Whilst arranging the set up for the first variable, the researcher casually said to the participant either (for the high and low reliability conditions):

*"I was discussing this with a [science or physics teacher/nursery child] last week. They said that they thought that the weight of the car is does not make a difference to how far the car travels. What do you think?"*

Or (for the no information condition):

*"So what effect do you think, say, weight will have on how far the car travels"*

### Test phase II

After this, the participant was asked for their prediction regarding the distance the car would travel, how sure they were, and why did they think that, as before.

### Intervention

The participant was told by the researcher they could now have another go, but this time it would be different, because the researcher was going to tell them which setups to try. The researcher said the following:

*"We are going to do a fair test. Do you know what that is?"*

After the participant responded, the researcher continued:

*"That is when you change only one thing, but leave everything else the same."*

This was said regardless of the participant's response.

*"You are going to do a fair test on weight. I want you to do this for the light car, medium car, and the heavy car. You are going to do this a couple of times. You are always going to do this from the same place."*

The researcher then indicated where the participant should do the fair test from (in the standard set up position). The participant then conducted the fair test by running the car down the track with each of the three weights, and then repeating it. They could do the fair test in any order they wished. The apparatus was fairly reliable, and the car almost always landed in box number 4, regardless of weight (there is a tiny difference with lighter cars travelling further than heavier cars, although this is rarely noticeable under these conditions). The researcher checked the participant was doing it correctly, and made sure the car was straight on the incline. If the car was not straight, it would bang into the side and not land in box number 4. If this clearly happened, the researcher commented that it has happened, and told the participant they could have another go.

TEST PHASE III

Once the participant has finished the fair test, the researcher reminded them of the standard set up and how to use the stick scale (if needed), and then the participant was asked for their prediction regarding the distance the car would travel under each set up, how sure they were, and why did they think that, as before.

Finally, to establish that the participant did, in fact, think that science teachers are more reliable than nursery children, they were asked:

*"Out of 10, how often do you think a [nursery child; science or physics teacher] would be right if you ask them a question?"*

## 2.8 ANALYSIS

### PRACTICE TRIALS

A binary variable was created for each practice trial measure (sequential fair trials, extreme set up, repeats). If the participant did at least one trial of the measured kind, they were coded as 1, otherwise they were coded as 0. This then created two groups that could be assessed according to age, BPVS score and gender. Between-participant t-tests were used for continuous variables and chi-square tests for categorical variables.

### WEIGHT PREDICTION AND EXPLANATION

The aim of the analysis was to examine the impact of hearing new information regarding the effect of weight on how far a car travelled on causal reasoning. Of particular interest was whether reliability of the source of the new information affected participants' reasoning regarding the effect of weight.

To assess children's initial beliefs regarding the effect of weight at baseline, the frequency of each type of prediction (heavy travels further, light travels further, no difference) was calculated. Chi-square was used to examine the relationship between initial belief about weight and age group.

To check whether the participants did think science or physics teachers were more reliable than nursery children, and whether this was impacted by participant age, a two-way mixed model ANOVA was performed on the ratings data.

For weight prediction, a binary variable was created where participants who made a correct weight prediction were coded as 1, and participants who made an incorrect prediction were coded as 0.

A similar binary variable was also created for weight explanation, where correct explanations were coded as 1, and incorrect explanations as 0. They would be considered to have made an incorrect explanation if they scored 0 - 1. They would be considered to have made a correct score if they scored 3 or above (see section 2.6.2 for more details regarding score criteria). Coding participants who scored 2 (ambiguous explanations) was more complex as it was not clear that 'ambiguous' explanations should be included in the incorrect *or* correct explanation category.

Here they could have been trying and failing to make an incorrect explanation *or* trying and failing to make a correct explanation. Given that the participants were generally capable of making explanations that explained their actual predictions regarding weight when the explanation concurred with their beliefs, it was possible that hearing new information that challenges one's beliefs regarding the effect of weight may have led to a more confused explanation, especially as the participants did not have much time between hearing the new information and providing an explanation. If this was the case, one would expect participants to be particularly confused in the high reliability condition where they have heard information incompatible with their beliefs from a source that they would usually trust. Examples of these types of explanation can be seen in Table 2-7, and suggest that some did seem confused by the new information. Given this, and that the number making correct explanations were too small for suitable statistical analyses to be carried out, it was decided that 0-1 would be coded as incorrect, and 2-5 coded as correct (and attempted correct) explanations.

Univariate analyses were conducted on both the weight prediction and explanation data. Firstly, this was to assess whether source reliability had an impact on participants' reasoning, and secondly, to identify potential relationships between factors commonly known to influence causal reasoning. These analyses were done at each of three time points: at baseline; after hearing information regarding weight; and after witnessing that weight did not affect how far the car travelled. Chi-square tests were used for categorical variables and between-participant t-tests for continuous variables. The odds ratios are presented for participants who correctly responded to information regarding the effect of weight, relative to participants who did not correctly respond. Interrelationships between personal characteristics variables were also assessed, including a check for multicollinearity between age and BPVS, age and gender, and gender and BPVS.

However, some of the explanatory variables were interrelated. In order to examine which factors had an independent influence on responses regarding the effect of weight, multivariate analyses using logistic progression programmes from IBM SPSS Version 23, suitable for estimation influences on binary outcome variables were performed. Multiple logistic regression was used to examine which factors had independent effects on making correct predictions regarding weight. Dummy

coding was used to enter source reliability into the analysis, where the no information group was used as a reference category. In order to facilitate comparison across the three time points, the source reliability and degree of conviction variables (our primary theoretical focus) were included in all the logistic regression models. However, of the other variables, only those that proved significant in the univariate analyses (age, gender and BPVS score) were entered into the logistic regression models at each time point.

### CAUSAL VARIABLES

Exploratory analyses were undertaken on the causal variable prediction date to examine participants' understanding of the causal variables, and how that might change after receiving information regarding one of the variables (weight) in the system.

Correct causal variable predictions were coded as 1, and incorrect causal variable predictions (both wrong direction, and same distance predictions) as 0. Changes in the number of correct predictions over time, and between causal variables were assessed using McNemar's tests. Differences between size of causal variable prediction difference were compared using within-participant one-way ANOVA.

The univariate analyses reflect the number of participants who provided complete responses for each relevant response. The multivariate regression analyses include only participants who gave complete responses. All variables in the final models met the 5% level of significance. All P values are 2-tailed.

# 3 RESULTS

## 3.1 PRACTICE TRIALS

### NUMBER OF REPEAT TRIALS

Participants repeated at least one practice trial 30% of the time. There was an effect of age where participants who repeated at least one trial were significantly younger, $t(158) = 4.04$, $p<.001$, (see Table 3-1).

Table 3-1 Number of participants (%), mean age (S.D.) and mean BPVS Score (S.D.) in relation to repeating trials, using extreme set up and doing a sequential fair test (N=160).

| | N (%) | Mean Age (S.D.) | BPVS[1] Score (S.D.) |
|---|---|---|---|
| **Repeat Trials** | | | |
| Repeat Trials | 48 (30%) | 9.95 (3.40) | 103.63 (16.95) |
| No Repeat Trials | 112 (70%) | 12.29 (3.33) | 99.80 (17.66) |
| **Extreme Set Up** | | | |
| Extreme Set Up | 67 (42%) | 11.31 (3.14) | 105.15 (16.74) |
| No Extreme Set Up | 93 (58%) | 11.78 (3.75) | 97.92 (16.74) |
| **Sequential Fair Test** | | | |
| Fair Test | 94 (59%) | 11.53 (3.73) | 100.03 (18.27) |
| No Fair Test | 66 (41%) | 11.66 (3.19) | 102.26 (16.36) |

[1] British Picture Vocabulary Scale

Further analyses by age group showed that within the primary school age participants, those in the younger age groups were more likely to repeat trials, $\chi^2(5) = 16.88$, $p<.01$: 55% of 6-7 year olds, 40% of 8-9 year olds, and 27% of 10-11 year olds repeated trials. From the 12-13 age group onwards, there was no evidence of any difference by age (see Table 3-2).

BPVS scores were not associated with repeating trials; the mean BPVS score for participants who made repeats was similar to those who did not, $t(158) = 1.23$, $p>.05$ (see Table 3-1).

There was a trend for more males to make more repeat trials than females (38% and 26% respectively; see Table 3-2). but this difference was not significant, $\chi^2(1) = 2.62$, $p=.11$

Table 3-2 Frequency of repeat trials (%) by age group and gender (N = 156).

| | | Number of Repeats (%) |
|---|---|---|
| **Age Group** | | |
| 6-7 years | 33 | 18 (55%) |
| 8-9 Years | 25 | 10 (40%) |
| 10-11 Years | 26 | 7 (27%) |
| 12-13 Years | 18 | 3 (17%) |
| 14-15 Years | 28 | 5 (18%) |
| 16-17 Years | 26 | 5 (19%) |
| | | |
| **Gender** | | |
| Female | 96 | 25 (26%) |
| Male | 60 | 23 (38%) |

EXTREME SET UP TRIALS

At least one extreme set up was used by 42% of the participants as one of their practice trials. Age was not related to participants choosing to do practice trials using set ups that would indicate the furthest and least furthest the car could travel, $t(158) = 0.85$, $p>.05$ (see Table 3.1).

However, participants conducting an extreme set up practice trial were more likely to have higher BPVS scores than those who did not, $t(158) = 2.63$, $p<.01$(see Table 3-1).

There was also an effect of gender, whereby males were more likely to use an extreme set up than females ($\chi^2(1)=6.79$, $p<.01$; see Table 3-3;).

Table 3-3 Frequency of participants (%) who used an extreme set up as one of their trials, or did a sequential fair test, by gender (N = 160).

| | N | Extreme Set Up (%) | Sequential Fair Tests (%) |
|---|---|---|---|
| **Gender** | | | |
| Females | 100 | 34 (34%) | 58 (58%). |
| Males | 60 | 33 (55%) | 36 (60%) |

### SEQUENTIAL FAIR TEST TRIALS

A sequential fair test on one of the three causal variables or weight was done by 59% of participants. Age was not related to whether participants did a sequential fair test, $t(158) = .23$, $p>.05$. Similarly, BPVS scores were not associated with doing a sequential fair test, $t(158) = 0.79$, $p>.05$ (see Table 3-1). Neither was there any relationship with gender, $\chi^2(1) = .06$, $p>.05$ (see Table 3-3).

## 3.2 PERSONAL CHARACTERISTICS INTER-RELATIONSHIPS

There is a statistically significant negative relationship between age and BPVS score, whereby older participants were likely to have lower BPVS scores, $r(160) = -.57$, $p<.001$. The variance inflation factor (VIF) was calculated to assess multicollinearity between them, VIF=1.15. As the VIF value was not substantially greater than 1, multicollinearity was not considered an issue (Bowerman & O'Connell, 1990).

There were also significant relationships between age and gender in that females were on average older, $t(158) = 5.28$, $p<.001$, VIF = 1.16, and between BPVS scores and gender, with males on average having higher scores, $t(158) = 4.78$, $p<.001$, VIF = 1.19 (see Table 3-4). Multicollinearity was not considered an issue in either case.

Table 3-4 Mean age (S.D.) in years and BPVS Score (S.D.) by gender (N = 160).

| | N | Mean Age (S.D.) | BPVS Score (S.D.) |
|---|---|---|---|
| **Gender** | | | |
| Female | 100 | 12.63 (3.50) | 96.15 (16.25) |
| Male | 60 | 9.84 (2.75) | 108.85 (16.63) |

## 3.3 PARTICIPANTS' INITIAL BELIEFS REGARDING THE EFFECT OF WEIGHT

Almost half (45%) of participants believed the light car would travel further while a similar proportion (41%) of participants believed the heavier car would travel further, with only a minority (14%) expecting no difference. This was related to age group, where younger participants were more likely to predict that the heavier car

would travel further, and older participants were more likely to predict the lighter car would travel further, $\chi^2(10) = 20.60$, $p<.05$ (see Table 3.5).

Table 3-5 Prediction of distance travelled based on car weight, by age group (N = 160), percentage of year group in parentheses.

| Age Group | Which car travels further? | | |
|---|---|---|---|
| | **Lighter (%)** | **No difference (%)** | **Heavier (%)** |
| 6-7 Years | 10 (30%) | 4 (12%) | 19 (38%) |
| 8-9 Years | 8 (32%) | 2 (8%) | 15 (60%) |
| 10-11 Years | 12 (46%) | 4 (15%) | 10 (39%) |
| Primary School | 30 (36%) | 10 (12%) | 44 (52%) |
| 12-13 Years | 6 (33%) | 2 (11%) | 10 (56%) |
| 14-15 Years | 19 (59%) | 7 (22%) | 6 (19%) |
| 16-17 Years | 17 (65%) | 3 (12%) | 6 (23%) |
| Secondary School | 42 (55%) | 12 (16%) | 22 (29%) |

## 3.4 ASSESSMENT OF SOURCE RELIABILITY MANIPULATION

To establish that the source reliability manipulation included appropriately different informants, participants were asked on a scale of 1 to 10, how likely they thought it was that a science/physics teacher, and nursery child, would be right if you asked them a question. As expected, 98% of participants rated the science/physics teacher as more reliable than a nursery child.

A 2 (Reliable Source: nursery child, science/physics teacher) X 6 (Age Group) Mixed Model ANOVA was performed on the reliability ratings data, where age group was a between-participant variable, and reliable source a within-participant variable. Participants rated the science/physics teacher as a more reliable source than the nursery child, $F(1,154) = 770.43$, $p<0.001$ (see Table 3-6).

There was also an effect of age $F(5,154) = 2.52$, $p < .05$, where younger participants gave higher ratings overall than older participants. However, pairwise comparisons showed no significant differences between age groups.

There was a significant interaction between age and reliability rating, $F(5, 154) = 5.87$, $p<.001$. All age groups made similar predictions regarding the reliability of a nursery child, whereas older participants rated science/physics teachers as less reliable sources compared with younger participants (see Figure 3-1).

Table 3-6 Mean participant ratings for the science teacher (S.D.)/nursery child (S.D.) by age group.

| | **Ratings** | |
|---|---|---|
| **Age Group** | **Science Teacher (S.D.)** | **Nursery Child (S.D.)** |
| 6-7 Years | 9.72 (.57) | 3.69 (1.98) |
| 8-9 Years | 9.22 (.96) | 4.04 (1.40) |
| 10-11 Years | 8.69 (1.05) | 4.42 (1.60) |
| 12-13 Years | 7.88 (1.64) | 4.00 (1.24) |
| 14-15 Years | 7.93 (1.70) | 4.38 (1.29) |
| 16-17 Years | 8.42 (1.58) | 3.92 (1.35) |

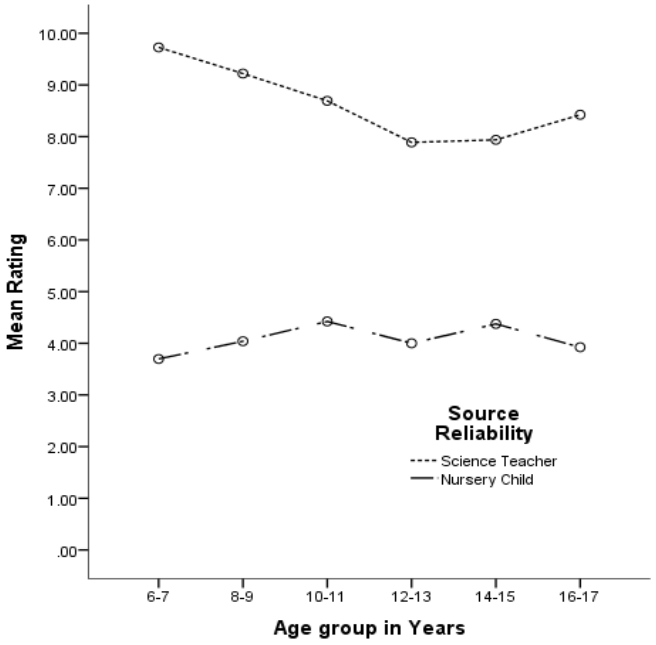


Figure 3-1 Mean rating of reliability for science/physics teacher and nursery child by age group.

A post hoc between-participant one-way ANOVA was carried out on both the science/physics and nursery child ratings data comparing age groups. Ratings of the nursery child's reliability showed no differences between ages, $F(5, 154) = .98$, $p>.05$. However, the science/physics teacher ratings showed an effect of age, $F(5. 154) = 8.83$, $p<.001$. Pairwise comparisons on age group showed that younger participants had a higher score for science/physics teacher ratings than older participants (see Table 3-7).

Table 3-7 Mean reliability rating (S.D.) and significant pairwise comparisons between age groups for the science/physics teacher and nursery child (N = 160).

| Age Group | N | Mean Teacher Rating (S.D.) | Pairwise Comparisons for Science/Physics Teacher | Mean Nursery Child Rating (S.D.) | Pairwise Comparisons for Nursery Child |
|---|---|---|---|---|---|
| 6-7 | 33 | 9.7 (0.6) | >10-11, 12-13, 14-15, 16-17 | 3.7 (2.0) | n.s. |
| 8-9 | 25 | 9.2 (1.0) | >12-13, 14-15 | 4.0 (1.4) | n.s. |
| 10-11 | 26 | 8.7 (1.0) | n.s. | 4.4 (1.6) | n.s. |
| 12-13 | 18 | 7.9 (1.6) | n.s. | 4.0 (1.2) | n.s. |
| 14-15 | 32 | 7.9 (1.7) | n.s. | 4.4 (1.3) | n.s. |
| 16-17 | 26 | 8.4 (1.6) | n.s. | 3.9 (1.4) | n.s. |

## 3.5 WEIGHT PREDICTION

### 3.5.1 UNIVARIATE ANALYSES FOR WEIGHT PREDICTION:

#### 3.5.1.1 SOURCE RELIABILITY

At baseline, before participants received new information regarding the causal system from differentially reliable sources, there was no relationship between source reliability and participants who made correct predictions regarding weight and those who did not (see Table 3-8).

After hearing new information regarding the causal system, participants whose new information came from a high reliability source were more likely to make a correct response, compared with participants who were told that the information was from a low reliability source, or participants who heard no new information, $\chi 2(2)=7.99$, $p<.05$ (see Table 3-8).

After participants had intervened personally on the causal system and witnessed that weight did not affect how far the car travelled, most participants made the correct response regarding the effect of weight, regardless of source reliability. There was no relationship between source reliability and participants who made correct responses or not (see Table 3-8).

#### 3.5.1.2 Degree of conviction

At baseline, participants who made a correct prediction were no more convinced by their prediction than participants who did not make a correct prediction. Also, there were no differences related to source reliability (see Table 3-9).

After hearing new information from differentially reliable sources, participants who made correct predictions were more convinced by their prediction than those who made incorrect predictions, $F(2,154)=3.78$, $p<.05$. This difference was related to source reliability whereby participants who made correct predictions in the high reliability condition had on average a higher level of conviction about their prediction than participants who made incorrect predictions ($p<.001$). This was not the case for participants in the low reliability and no information conditions (see Table 3-9).

After witnessing that weight did not affect distance travelled, participants who made correct predictions had on average a higher level of conviction about their response compared with participants who made incorrect predictions regarding weight, $F(2,154)=7.31$, $p<.01$. This difference was also related to source reliability. Participants who made a correct prediction had on average more conviction, compared with those who made incorrect predictions, in both the low reliability ($p<.001$) and no information conditions ($p<.001$). This was not the case for participants in the high reliability condition (see Table 3-9).

Table 3-8 Univariate analyses of the relevance of source reliability and gender (categorical factors) at each time point for weight prediction (N = 160).

| | | Baseline | | Post Information | | Post Intervention | |
|---|---|---|---|---|---|---|---|
| **Factor** | **N** | **Correct Prediction (%)** | **Odds Ratio (95% C.I.)** | **Correct Prediction (%)** | **Odds Ratio (95% C.I.)** | **Correct Prediction (%)** | **Odds Ratio (95% C.I.)** |
| **Source Reliability** | | | | | | | |
| High reliability Source | 55 | 7 (13%) | 0.88 (.33-2.29) | 21 (38%) | 2.80 (1.34-5.84)** | 50 (91%) | 1.17 (.39-3.56) |
| Low reliability Source | 55 | 6 (11%) | 0.68 (.25-1.85) | 11 (20%) | 0.66 (.30-1.44) | 51 (93%) | 1.64 (.51-5.37) |
| No Information | 50 | 9 (16%) | 1 | 8 (15%) | 1 | 43 (86%) | 1 |
| | | $\chi2(2)=1.18$, p>.05 | | $\chi2(2)=7.99$, p<.05 | | $\chi2(2)=1.39$, p>.05 | |
| **Gender** | | | | | | | |
| Male | 60 | 10 (17%) | 1.47 (.59-3.64) | 21 (35%) | 2.30 (1.11-4.76)* | 54 (90%) | 1.00 (.34-2.91) |
| Female | 100 | 12 (12%) | 1 | 19 (19%) | 1 | 90 (90%) | 1 |
| | | $\chi2(1)=.69$, p>.05 | | $\chi2(1)=5.12$, p<.05 | | $\chi2(1)=0$, p>.05 | |

*P<.05,**P<.01

Table 3-9 Univariate analyses of age, BPVS, and degree of conviction (continuous factors) at each time point for the weight prediction outcome.

| | Baseline | | | | Post Info | | | | Post Intervention | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Factor** | **n** | **Mean (S.D.)** | **T value** | **P** | **n** | **Mean (S.D.)** | **T value** | **P** | **n** | **Mean (S.D.)** | **T value** | **P** |
| **Age** | | | | | | | | | | | | |
| Correct Prediction | 22 | 12.18 (3.62) | 0.85 | n.s | 40 | 11.18 (3.30) | -0.835 | n.s | 144 | 11.86 (3.44) | 3.12 | <.01 |
| Incorrect Prediction | 138 | 11.49 (3.49) | | | 120 | 11.72 (3.57) | | | 16 | 9.06 (3.16) | | |
| **BPVS** | | | | | | | | | | | | |
| Correct Prediction | 22 | 102.36 (19.03) | 0.41 | n.s | 40 | 104.63 (17.07) | 0.99 | n.s | 144 | 101.31 (17.39) | 0.79 | n.s |
| Incorrect Prediction | 138 | 100.72 (17.29) | | | 120 | 99.73 (17.52) | | | 16 | 97.69 (18.56) | | |
| **Degree of Conviction** | | | | | | | | | | | | |
| Correct Prediction | 22 | 9.91 (2.04) | 0.13 | n.s | 40 | 11.23 (2.21) | 4.53 | <.001 | 144 | 13.25 (1.38) | 6.53 | <.001 |
| Incorrect Prediction | 138 | 9.85 (2.07) | | | 120 | 9.53 (2.00) | | | 16 | 10.56 (2.73) | | |
| **Degree of Conviction - Interaction Effects with Source Reliability** | | | | | | | | | | | | |
| **High Reliability** | | | | | | | | | | | | |
| Correct Prediction | 7 | 10.00 (2.38) | -0.07 | n.s | 21 | 11.95 (1.77) | 5.56 | <.001 | 50 | 13.22 (1.47) | 0.89 | n.s |
| Incorrect Prediction | 48 | 10.06 (2.19) | | | 34 | 9.15 (1.84) | | | 5 | 12.6 (1.67) | | |
| **Low Reliability** | | | | | | | | | | | | |
| Correct Prediction | 6 | 9.67 (2.66) | -0.06 | n.s | 11 | 9.91 (2.77) | 0.58 | n.s | 51 | 13.29 (1.29) | 5.69 | <.001 |
| Incorrect Prediction | 49 | 9.71 (1.91) | | | 44 | 9.43 (2.36) | | | 4 | 8.75 (3.77) | | |
| **No Information** | | | | | | | | | | | | |
| Correct Prediction | 9 | 10.00 (1.50) | 0.33 | n.s | 8 | 11.13 (1.73) | 1.85 | n.s | 43 | 13.23 (1.41) | 5.09 | <.001 |
| Incorrect Prediction | 41 | 9.76 (2.13) | | | 42 | 9.93 (1.67) | | | 7 | 10.14 (1.95) | | |
| | | | F(2,154)=.05, p>.05 | | | | F(2,154)=3.78, p<.05 | | | | F(2,154)=7.31, p<.01 | |

Table 3-10 Results of a logistic regression to predict a correct prediction regarding the effect of weight on distance travelled.

| **Explanatory Variables** | **Baseline** | | | **Post Information** | | | **Post Intervention** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Odds ratio (95% CI)** | **P** | **Wald** | **Odds ratio (95% CI)** | **P** | **Wald** | **Odds ratio (95% CI)** | **P** | **Wald** |
| High Reliability | 0.66 (.23-1.94) | n.s | 0.57 | 3.62 (1.32-9.93) | p<.05 | 6.25 | 0.85 (.19-3.73) | n.s | 0.05 |
| Low Reliability | 0.56 (.18-1.70) | n.s | 1.05 | 1.71 (.58-5.01) | n.s | 0.95 | 2.65 (.46-15.22) | n.s | 1.20 |
| Degree of Conviction | 1.02 (.81-1.27) | n.s | 0.02 | 1.54 (1.24-1.91) | p<.001 | 15.30 | 2.14 (1.51-3.03) | p<.001 | 18.14 |
| Gender | Not in model | | | 2.76 (1.21-6.30) | p<.05 | 5.84 | Not in model | | |
| Age | Not in model | | | Not in model | | | 1.37 (1.12-1.69) | p<.01 | 9.02 |
| | $\chi^2(3) = 1.17$, p>.05 | | | $\chi^2(4) = 32.65$, p<.001 | | | $\chi^2(4) = 37.50$, p<.001 | | |

#### 3.5.1.3 Personal characteristics

##### Age

After participants had observed that weight does not affect how far the car travels, participants who made correct predictions regarding the effect of weight were likely to be older than participants who made incorrect predictions, $t(158) = 3.12$, $p<.01$. There were no differences in age relating to correctness of predictions at baseline or after hearing new information (see Table 3-9).

##### Gender

There was no relationship between gender and number of correct predictions at baseline. However, boys were more likely than girls to make correct predictions after hearing that weight does not affect distance travelled, $\chi2(1) = 5.12$, $p<.05$. This relationship was not evident when participants observed that weight did not affect distance travelled, where the majority of participants made correct predictions regarding weight different (see Table 3-8).

##### Language

There was no significant association between receptive language based on BPVS scores and making a correct prediction at any time point (see Table 3-9).

### 3.5.2 Multivariate analyses of correct predictions regarding the effect of weight

The factors found by use of logistic regression to be independently associated with making a correct prediction regarding the effect of weight at each time point, once other factors had been taken into account are presented in Table 3-10.

At baseline, the included predictors, the high and low reliability variables, as well as degree of conviction did not function as accurate predictors of a correct prediction, $\chi^2(3) = 1.17$, $p>.05$.

After participants heard new information, the prediction accuracy of the model was 82%, $\chi^2(4) = 32.65$, $p<.001$. Participants who heard new information regarding the effect of weight from a high reliability source were 4.36 times more likely to make a correct prediction, compared with participants who heard no new information, whereas hearing information from a low reliability source was not an independent

predictor of correctness. Furthermore, male participants were 2.76 times more likely to make correct predictions compared with females. Degree of conviction in participants' response was also associated with a greater likelihood of making a correct explanation (OR = 1.54). Neither age nor hearing information from a low reliability source were significant independent predictors of correctness (see Table 3-10).

After participants had intervened on the system, the prediction accuracy of the model was 94%, $\chi^2(4) = 37.50$, $p<.001$. Hearing new information from either a high or low reliability source did not improve the odds of making correct predictions, compared with hearing no information. However, degree of conviction continued to function as an independent predictor, where more conviction increased the odds of making a correct prediction (OR = 2.14). Likewise, age was also an independent predictor, where older participants were more likely to make correct predictions regarding the effect of weight (OR = 1.37; see Table 3-10).

## 3.6 WEIGHT EXPLANATION

As the recordings of the verbal explanations were sometimes inaudible due to both external noise outside the testing room, and quietness of the voice of the participant, there are missing data points. One child contributed no data points due to inaudibility, and 6 data points were missing at baseline, 0 post reliable information, and 2 post intervention.

The majority of participants provided explanations that were incorrect (scoring 1) at baseline (89%) and after hearing information regarding the effect of weight (78%). After participants witnessed that weight did not affect how far the car travelled, the number of incorrect predictions (scoring 1) decreased to 23% (see Table 3-11). Very few participants did not provide an explanation for their weight prediction (scoring 0) - 3% at baseline and 2% after hearing new information. The proportion was higher after participants witnessed that weight did not affect distance travelled (54%).

Few participants made explanations scoring three or more: 3% at baseline, 9% after hearing new information, and 12% after witnessing that weight did not affect distance travelled.

Considering ambiguous explanations, only 5% of participants scored 2 at baseline. However, after hearing new information, the proportion of participants making ambiguous explanations appeared to be higher (11%). Furthermore, to examine whether being in the high reliability condition led to an increase in ambiguous explanations, post information participants in the high reliability condition achieved both more correct explanation scores and more ambiguous scores than the low reliability or no information groups after hearing that weight does not affect distance travelled, $\chi^2(6) = 13.64$, $p<.05$ (see Table 3-11).

Table 3-11 Quality of explanations at each time point and for each source reliability condition, after hearing that weight does not affect distance travelled.

| | **Explanation Score (%)** | | | | |
|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3+** | **Total** |
| **Time of Testing** | | | | | |
| Baseline | 5 (3%) | 136 (89%) | 7 (5%) | 4 (3%) | 153 |
| Post Information | 3 (2%) | 122 (78%) | 18 (11%) | 14 (9%) | 157 |
| Post Intervention | 87 (56%) | 36 (23%) | 14 (9%) | 19 (12%) | 156 |
| **Source Reliability (Post Information, N=157)** | | | | | |
| High Reliability | 1 (2%) | 36 (65%) | 9 (16%) | 9 (16%) | 55 |
| Low Reliability | 2 (4%) | 42 (79%) | 4 (8%) | 5 (8%) | 53 |
| No Information | 0 (0%) | 44 (90%) | 5 (10%) | 0 (0%) | 49 |
| | $\chi^2(6)=13.64$, $p<.05$ | | | | |

### 3.6.1 Univariate analyses for weight explanation

#### 3.6.1.1 Source reliability

At baseline, before participants received information regarding the causal system from differentially reliable sources, there was no relationship between source reliability and participants who made correct explanations regarding weight and those that did not (see Table 3-12).

After hearing new information regarding the causal system, participants who heard new information from a high reliability source were more likely than expected to make a correct explanation (33%), compared with participants who were heard new information from a low reliability source (17%), or participants who heard no new information (10%), $\chi2(2) = 8.67$, $p<.05$ (see Table 3-12).

After participants intervened on the causal system, and witnessed that weight did not affect how far the car travelled, source reliability again was not related significantly to correctness of participants' explanations (see Table 3-12).

#### 3.6.1.2 Degree of conviction

After the 'receiving information' stage (post information) participants who made correct explanations had on average higher scores for degree of conviction about explanations than participants who did not make correct explanations, $t(158) = 3.29$, $p<.01$. However, there was no significant difference between conviction scores as they related to correctness of explanation at baseline or after participants had witnessed for themselves that weight did not affect how far the car travelled (see Table 3-13).

#### 3.6.1.3 Personal characteristics

##### Age

Correctness of explanations was not related to age, at any time point (see Table 3-13).

##### Gender

There was no relationship between gender and making a correct explanation at baseline. However, a greater proportion of boys made correct explanations after hearing that weight does not affect distance travelled, compared with girls, $\chi2(1)=4.14$, $p<.05$. This relationship was not evident after witnessing that weight did not affect distance travelled, where the proportion of participants who made correct explanations was similar by gender (see Table 3-12).

##### Language

As with weight prediction, making a correct explanation was not related to the BPVS score at any time point (see Table 3-13).

Table 3-12 Univariate analyses of source reliability and gender (categorical factors) at each time point for making a correct weight explanation.

| | Baseline | | | Post Information | | | Post Intervention | | |
|---|---|---|---|---|---|---|---|---|---|
| **Factor** | **N** | **Correct Explanation (%)** | **Odds Ratio (95% C.I.)** | **N** | **Correct Explanation (%)** | **Odds Ratio (95% C.I.)** | **N** | **Correct Explanation (%)** | **Odds Ratio (95% C.I.)** |
| Total | 153 | | | 157 | | | 156 | | |
| **Source Reliability** | | | | | | | | | |
| High reliability Source | 53 | 4 (6%) | 0.94 (.27-3.27) | 55 | 18 (33%) | 3.06 (1.38-6.79)** | 55 | 16 (29%) | 2.03 (.93-4.43) |
| Low reliability Source | 53 | 4 (6%) | 0.94 (.27-3.27) | 53 | 9 (17%) | 0.72 (.31-1.69) | 53 | 7 (13%) | 0.45 (.18-1.12) |
| No Information | 47 | 4 (9%) | 1 | 49 | 5 (10%) | 1 | 48 | 10 (21%) | 1 |
| | | $\chi2(2)=0.04$, $p>.05$ | | | $\chi2(2)=8.67$, $p<.05$ | | | $\chi2(2)=4.09$, $p>.05$ | |
| **Gender** | | | | | | | | | |
| Male | 56 | 4 (7%) | 0.87 (.25-2.98) | 59 | 17 (29%) | 2.24 (1.02-4.92)* | 60 | 11 (18%) | 0.76 (.34-1.70) |
| Female | 97 | 8 (8%) | 1 | 98 | 15 (15%) | 1 | 96 | 22 (23%) | 1 |
| | | $\chi2(1)=0.06$, $p>.05$ | | | $\chi2(1)=4.14$, $p<.05$ | | | $\chi2(1)=0.47$, $p>.05$ | |

*P<.05,**P<.01

Table 3-13 Univariate analyses of age, BPVS, and degree of conviction (continuous factors) at each time point for making a correct weight explanation.

| | Baseline | | | | Post Information | | | | Post Intervention | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Factor** | **n** | **Mean (S.D.)** | **T value** | **P** | **n** | **Mean (S.D.)** | **T value** | **P** | **n** | **Mean (S.D.)** | **T value** | **P** |
| Total | 153 | | | | 157 | | | | 156 | | | |
| **Age** | | | | | | | | | | | | |
| Correct Explanation | 12 | 12.88 (4.26) | 1.23 | n.s | 32 | 11.47 (3.28) | -0.28 | n.s | 33 | 11.33 (3.12) | -0.61 | n.s |
| Incorrect Explanation | 141 | 11.57 (3.48) | | | 125 | 11.67 (3.59) | | | 123 | 11.75 (3.61) | | |
| **BPVS** | | | | | | | | | | | | |
| Correct Explanation | 12 | 99.92 (20.72) | -0.18 | n.s | 32 | 104.03 (18.73) | 1.14 | n.s | 33 | 102.12 (19.89) | 0.43 | n.s |
| Incorrect Explanation | 141 | 100.87 (17.35) | | | 125 | 100.07 (17.22) | | | 123 | 100.63 (17.06) | | |
| **Degree of Conviction** | | | | | | | | | | | | |
| Correct Explanation | 12 | 10.83 (1.80) | 1.73 | n.s | 32 | 11.06 (2.17) | 3.29 | p<.01 | 33 | 13.45 (1.12) | 1.74 | n.s |
| Incorrect Explanation | 141 | 9.77 (2.06) | | | 125 | 9.69 (2.30) | | | 123 | 12.85 (1.89) | | |

Table 3-14 Results of a logistic regression to predict a correct explanation regarding the effect of weight on distance travelled.

| **Explanatory Variables** | **Baseline** | | | **Post Information** | | | **Post Intervention** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Odds ratio (95% C.I.)** | **P** | **Wald** | **Odds ratio (95% C.I.)** | **P** | **Wald** | **Odds ratio (95% C.I.)** | **P** | **Wald** |
| High Reliability | 0.78 (.18-3.41) | n.s. | 0.11 | 4.70 (1.51-14.59) | p<.01 | 7.17 | 1.46 (.58-3.66) | n.s. | 0.65 |
| Low Reliability | 0.94 (0.22-4.05) | n.s. | 0.01 | 2.40 (.71-8.12) | n.s. | 1.98 | 0.55 (.19-1.59) | n.s. | 1.22 |
| Degree of Conviction | 1.34 (.96-1.87) | n.s. | 2.93 | 1.38 (1.12-1.70) | p<.01 | 8.81 | 1.30 (.96-1.76) | n.s. | 2.79 |
| Gender | Not in model | | | 2.64 (1.11-6.24) | p<.05 | 4.85 | Not in model | | |
| | $\chi^2(3)=0.11$, p>.05 | | | $\chi^2(3)=13.64$, p<.01 | | | $\chi^2(3)=7.61$, p>.05 | | |

### 3.6.2 MULTIVARIATE ANALYSES OF CORRECT EXPLANATION REGARDING THE EFFECT OF WEIGHT

At baseline, the included predictors - the high and low reliability variables, as well as degree of conviction did not function as significant predictors of a correct explanation, $\chi^2(3) = 0.11$, $p>.05$ (See Table 3-14).

After participants heard new information, the prediction accuracy of the model was 80%, $\chi^2(3) = 13.64$, $p<.01$. Participants who heard new information regarding the effect of weight from a high reliability source were 4.7 times more likely to make a correct explanation, compared with hearing no new information. Hearing information from a low reliability source was not an independent predictor of correctness. Gender was also related to the likelihood of making a correct explanation, whereby males were 2.64 times more likely to make a correct prediction, compared with females. The degree of conviction in their response also remained significantly related to making a correct explanation, where more conviction was associated with a greater likelihood of making a correct explanation (OR = 1.38; see Table 3-14).

After participants had witnessed that weight did not affect distance travelled, the model was not significant in relation to making a correct explanation, $\chi^2(3) = 7.61$, $p>.05$.

## 3.7 CAUSAL VARIABLE PREDICTION

#### CORRECT PREDICTIONS

At baseline, nearly all participants made correct predictions for the effect of the height of the incline (99%), and most made correct predictions about the relevance of the starting point on the incline (88%), and surface friction of the incline (86%). This pattern remained similar after participants had received new information regarding weight (post information; height – 98%; starting point – 91%; surface friction 88%). However, after witnessing that weight did not affect how far the car travelled (post intervention), the proportion of correct predictions decreased for height (89%), starting point (83%), and surface friction (68%; see Table 3-15).

### INCORRECT PREDICTIONS (NO EFFECT)

At baseline, no participants predicted that there was no effect of height of the incline on distance travelled (0%), where the prediction difference was 0. The proportion increased for starting point (8%), and more so for surface friction (11%). This pattern was similar post information (height –1%; starting point – 5%; surface friction - 11%). However, post intervention, the number of participants who predicted no effect increased for all three variables (height – 10%; starting point – 14%; surface friction - 31%; see Table 3-15).

### INCORRECT PREDICTIONS (WRONG DIRECTION)

Relatively few participants made prediction errors in the wrong direction, by predicting that a car on the lower set up for a particular causal variable would travel further than the car on the higher set up, with little difference across time points (height – 1%; starting point – 3-4%; surface friction – 1-3%; see Table 3-15).

In total, only 9% of participants made causal variable predictions in the wrong direction. For each causal variable, participants could make a prediction error at each of the three time points, for a total of three possible errors. Participants ranged from making predictions in the wrong direction once, to making errors at all three time points (see Table 3-16).

There was little difference between age groups (7-12%). The majority of errors related to starting point on the incline, where 80% of the participants who made errors made them regarding starting point, 33% for surface friction and 12% for height of incline. Three participants made prediction errors regarding two causal variables, and no participants made such errors for all three variables (see Table 3-16).

As the number of participants who incorrectly predicted no effect, was much higher than for participants who made prediction errors in the wrong direction, the two types of prediction errors were collapsed for analyses, whereby number of correct and incorrect predictions were compared.

Table 3-15 Number and percentage of participants making correct and incorrect predictions (no effect and wrong direction), and mean difference (S.D.) between high and low set up predictions at each time of testing.

| | Baseline | | Post Information | | Post Intervention | |
|---|---|---|---|---|---|---|
| **Factor** | **N(%)** | **Mean (S.D.)** | **N(%)** | **Mean (S.D.)** | **N(%)** | **Mean (S.D.)** |
| **All Cases** | **160** | | | | | |
| **Height of Incline** | | | | | | |
| Correct Prediction (steeper further) | 158 (99%) | 2.36 (.78) | 156 (98%) | 2.15 (.85) | 142 (89%) | 1.89 (.73) |
| Incorrect Prediction (no effect) | 0 (0%) | 0 | 2 (1%) | 0 | 16 (10%) | 0 |
| Incorrect Prediction (flatter further) | 2 (1%) | -2.00 (0) | 2 (1%) | -2.00 (1.41) | 2 (1%) | -1.50 (.71) |
| Incorrect Prediction Total | 2 (1%) | | 4 (2%) | | 18 (11%) | |
| **Starting Point on Incline** | | | | | | |
| Correct Prediction (higher further) | 141 (88%) | 1.82 (.71) | 146 (91%) | 1.84 (.67) | 133 (83%) | 1.64 (.63) |
| Incorrect Prediction (no effect) | 12 (8%) | 0 | 8 (5%) | 0 | 22 (14%) | 0 |
| Incorrect Prediction (lower further) | 7 (4%) | -1.86 (1.21) | 6 (4%) | -1.83 (.41) | 5 (3%) | -1.00 (0) |
| Incorrect Prediction Total | 19 (12%) | | 14 (9%) | | 27 (17%) | |
| **Surface Friction of Incline** | | | | | | |
| Correct Prediction (smoother further) | 137 (86%) | 1.66 (.65) | 141 (88%) | 1.62 (.69) | 108 (68%) | 1.56 (.60) |
| Incorrect Prediction (no effect) | 18 (11%) | 0 | 18 (11%) | 0 | 49 (31%) | 0 |
| Incorrect Prediction (rougher further) | 5 (3%) | -1.00 (0) | 1 (1%) | -2 | 3 (2%) | -1.00 (0) |
| Incorrect Prediction Total | 23 (14%) | | 19 (12%) | | 52 (33%) | |

Table 3-16 Age group, number and proportion of participants making incorrect predictions in the wrong direction. For each causal variable, there are three possible errors, one at each time point (number of errors/3).

| | N | No. of Predictions in Wrong Direction (%) | School | Causal Variables: Starting Point | Height | Surface Friction |
|---|---|---|---|---|---|---|
| **All Cases** | 160 | 15 (9%) | | | | |
| | | | | | | |
| **Age Group** | | | | | | |
| 6-7 years | 33 | 4 (12%) | A | 1/3 | 0 | 0 |
| | | | B | 3/3 | 0 | 0 |
| | | | B | 2/3 | 0 | 0 |
| | | | B | 1/3 | 0 | 0 |
| 8-9 Years | 25 | 2 (8%) | A | 1/3 | 0 | 0 |
| | | | B | 0 | 0 | 2/3 |
| 10-11 Years | 26 | 3 (12%) | B | 3/3 | 2/3 | 0 |
| | | | B | 0 | 0 | 1/3 |
| | | | B | 2/3 | 0 | 1/3 |
| 12-13 Years | 18 | 2 (11%) | C | 1/3 | 0 | 0 |
| | | | D | 0 | 0 | 2/3 |
| 14-15 Years | 28 | 2 (7%) | C | 1/3 | 0 | 0 |
| | | | C | 2/3 | 0 | 0 |
| 16-17 Years | 26 | 2 (8%) | C | 1/3 | 3/3 | 0 |
| | | | D | 0 | 0 | 2/3 |
| Total (%) | | | | 12 (80%) | 2 (13%) | 5 (33%) |

### 3.7.1 CAUSAL VARIABLE ANALYSIS

As observed earlier, the pattern of overall prediction errors changed across time of testing, whereby prediction errors remained relatively static across the first two time points, and increased after the third time point, most prominently for height (baseline – 1%; post information – 2%; post intervention – 11%) and surface friction (baseline – 14%; post information – 12%; post intervention – 33%). The pattern was less prominent for starting point (baseline – 12%; post information – 9%; post intervention – 17%; see Table 3-15). Given this, comparisons were made between baseline, and after participants had witnessed that weight did not affect distance travelled.

A McNemar's test compared the pattern of correct and incorrect predictions for each causal variable at baseline and after participants had witnessed that weight does not affect how far the car travels (post intervention).

For height of the incline, 89% of participants made a correct prediction both at baseline and post intervention. A further 10 % made correct predictions at baseline only, whereas no participants made correct predictions post intervention only ($p<.001$; see Table 3.16).

For starting point on the incline, 75% of participants made a correct prediction at baseline and post intervention. A further 13% made correct predictions at baseline only, and 4% made correct predictions post intervention only, ($p>.05$; see Table 3.16).

For surface friction on the incline, 62% of participants made correct predictions at baseline and post intervention. A further 24% made correct predictions at baseline only, whereas only 6% made correct predictions post intervention only ($p<.001$; see Table 3-17).

Table 3-17 No. of correct/incorrect predictions before and after the intervention on weight, and reports the outcome of McNemar's test, for each causal variable (N = 160).

| | **Height** | **Starting Point** | **Surface Friction** |
|---|---|---|---|
| Correct Predictions - *At baseline and Post Intervention* | 142 (89%) | 120 (75%) | 99 (62%) |
| Correct Predictions - *At baseline only* | 16 (10%) | 21 (13%) | 38 (24%) |
| Correct Predictions - *Post Intervention only* | 0 (0%) | 6 (4%) | 9 (6%) |
| Incorrect Predictions – *At baseline and post intervention* | 2 (1%) | 13 (8%) | 14 (9%) |
| P | p<.001 | n.s. | p<.001 |

There were more correct predictions for height than starting point or surface friction (at baseline: height – 99%; starting point – 88%; surface friction – 86%; post intervention: height – 89%; starting point – 83%; surface friction – 68%); where the proportion of correct predictions was greater for height at each time point. Surface friction saw the greatest number of prediction errors, where the number of correct

predictions was lowest at each time point, and particularly so after the weight intervention (see Table 3-15).

A McNemar's Test was used to assess the nature of this pattern between the three causal variables. As the number of prediction errors increased from height with the fewest errors, to starting point, then surface friction, height was compared with starting point, and starting point with friction, at each time of testing.

At baseline, 88% of participants made correct predictions regarding both height and starting point. An additional 11% of participants made correct predictions regarding height only, whereas only an additional 0.5% were correct regarding starting point ($p<.001$; see Table 3-18). For starting point compared with surface friction, 75% of participants made correct predictions regarding both variables. A similar proportion were correct regarding starting point only (13%) and surface friction only (11%; $p>.05$; see Table 3-18).

This pattern held after receiving new information (post information) for height and starting point (both correct – 91%), where participants made correct predictions regarding height only more frequently compared with starting point only (height - 7%; starting point - 0.5%; $p<.01$). For starting point and friction, the pattern was also similar (both correct – 82%), where starting point had a similar number of correct predictions (9%) compared with surface friction (6%; $p>.05$; see Table 3-18).

However, post intervention, the number of correct predictions overall for both height and starting point was 79%, and a similar proportion of participants were correct regarding height only (9%) and starting point only (4%; $p>.05$). For starting point and surface prediction, the number of correct predictions for both was 59%, where proportionally more were correct for starting point only (24%) compared with surface friction only (9%; see Table 3-18).

Table 3-18 Number of correct/incorrect predictions comparing height and starting point, and starting point and friction, at each time point, and reports the outcome of McNemar's test (N = 160).

| | Time of Testing | | |
|---|---|---|---|
| | **Baseline (%)** | **Post Information (%)** | **Post Intervention (%)** |
| Correct Prediction - *Height & Starting Point* | 140 (88%) | 145 (91%) | 127 (79%) |
| Correct Prediction - *Height Only* | 18 (11%) | 11 (7%) | 15 (9%) |
| Correct Prediction - *Starting Point Only* | 1 (.5%) | 1 (.5%) | 6 (4%) |
| Incorrect Prediction *Height & Starting Point* | 1 (.5%) | 3 (1.5%) | 12 (8%) |
| P | p<.001 | p<.01 | n.s. |
| Correct Prediction - *Starting Point & Surface Friction* | 120 (75%) | 131 (82%) | 94 (59%) |
| Correct Prediction - *Starting Point Only* | 21 (13%) | 15 (9%) | 39 (24%) |
| Correct Prediction – *Surface Friction Only* | 17 (11%) | 10 (6%) | 14 (9%) |
| Correct Prediction - *Starting Point & Surface Friction* | 2 (1%) | 4 (3%) | 13 (8%) |
| P | n.s. | n.s. | p<.001 |

As can be seen in Table 3-15, the size of the mean difference between predictions for the high and low set up for each causal variable differs, where the mean prediction difference for height was greatest, followed by starting point, and then friction. A within-participant one-way ANOVA was performed on the causal variable prediction difference data, comparing the three causal variables. There was a difference at baseline, $F(2, 318) = 7.34$, $p<.001$, where the height prediction difference was larger than the friction prediction difference ($p<.001$). After the weight intervention, the mean prediction difference ranking of height, starting point, friction was the same, $F(2, 318) = 18.33$, $p<.001$, where mean prediction difference for height was greater than starting point ($p<.001$) which was greater than friction ($p<.001$).

# 4 General Discussion

Source reliability plays a key role in judging the quality of information bearing on a particular issue, crucially important in both everyday life and scientific reasoning. Young children, adolescents and adults have all been shown to pay attention to source reliability. However, it has been difficult to understand whether there are developmental trajectories as the paradigms used to assess source reliability change substantially dependent on the age group that is being assessed. By studying a sample that spanned five to 17 years, using the same paradigm, this study aimed to contribute to the literature in the following ways: first, to establish whether participants would discriminate between differentially reliable sources when reasoning about a familiar causal system, and how this might change developmentally; second, to compare the developmental trajectories of implicit (prediction) and explicit (explanation) causal understanding; and third, to investigate whether gender and language are relevant in predicting children's reasoning regarding a familiar causal system when faced with unexpected information from differentially reliable sources. Additional goals were to conduct exploratory analyses examining how participants engaged in unguided 'play' with the causal system, and seeking to discover how well the causal system was understood.

## 4.1 Implicit understanding of the causal system (weight prediction)

It was predicted that participants would differentially utilise information from high and low reliability sources when reasoning about a familiar causal system. As expected, source reliability did appear to play a role in participants' reasoning. Participants who received unexpected, but 'true' information from a high reliability source were more likely to make a correct prediction regarding the effect of weight on distance travelled than participants who received no information. Participants who received information from a low reliability source were no more likely to make correct predictions regarding weight than participants who received no information. This concurs with the literature looking at source reliability, which suggests that humans are sensitive to source reliability information and use it to

inform reasoning regarding every day and scientific matters (e.g. Bråten et al., 2012; Hahn et al., 2005; Koenig & Harris, 2005).

### 4.1.1 Age

Contrary to expectation, age was not found to be related to making correct predictions regarding the effect of weight in this study, nor was it relevant when participants were faced with unexpected information from differentially reliable sources. It is possible that source reliability understanding occurs at a young age, and that the simplicity of the task requirements meant that there were therefore no age-related improvements in performance.

However, as predicted, age was relevant once participants had witnessed that weight did not affect distance travelled. At this point, most participants predicted that there would be no effect of the car's weight, but those who did incorrectly predict there would be an effect were more likely to be younger. It is possible that some younger children found it difficult to inhibit their strong prior beliefs regarding the causal system, even when faced with observational evidence that their prior beliefs were incorrect. Best et al. (2011) found evidence to suggest that performance on the three core executive functions, including inhibition, improved with age. Children experience significant improvements in performance in executive function tasks (including inhibition) over the ages of five to seven years. It might be that, for some of the study participants, the ability to demonstrate inhibition control was developing more slowly, resulting in their failing to adjust their predictions. There is evidence that children can revise their beliefs from as young as four to five years of age, as demonstrated by Macris and Sobel (2017) who found that participants revised their beliefs when faced with disconfirming evidence regarding an unfamiliar system (see also Koerber et al., 2005). Similarly, Schulz and Gopnik (2004) also found evidence to suggest that four- and five-year-olds were, in principle, capable of overriding prior beliefs regarding causal relations (physical causes lead to physical effects; psychological causes lead to psychological effects). However, the tasks were simple, and fairly artificial in construction, involving participants being asked to draw on prior beliefs regarding the world, in relation to puppets, and other toys. They were also asked to draw on their prior beliefs to make judgements about an unfamiliar causal system. It is possible that prior beliefs would

have had more of an impact in more realistic situations, and with more familiar systems. Even adult experts appear to have an intuitive response that they then need to inhibit, when faced with problems that do not align with prior beliefs about the world; they just happen to be better at inhibiting intuitive responses than novices (Masson, Potvin, Riopel, & Foisy, 2014). As such, it would be no surprise that the younger children also find it difficult.

An alternative explanation might be that some younger children did not fully assimilate the implications of what they had observed in this study. So rather than understanding the implication, but failing to inhibit an incorrect response, some younger participants may not have understood the implication of what they observed in relation to future predictions regarding the effect of the car's weight. However, since very young children can observe minimal evidence of the functioning of an *unfamiliar* causal system, and demonstrate understanding of the causal system via prediction (e.g. Gopnik et al., 2001), it seems unlikely that they lack the cognitive ability required to revise their beliefs regarding a *familiar* causal system (unless constrained by prior beliefs).

If some of the younger participants did not make appropriate predictions regarding weight following *observing* that weight does not affect distance the car travelled, then it is possible that they may be even less likely to override their prior beliefs when faced with second-hand information they have heard ("the science teacher told me…"). However, Fitneva (2008) found that nine-year-old participants preferred first-hand perceptually based evidence (information that was observed by the source), whereas, six-year-olds preferred cognitive sources (sources that claim to 'know' the information), and did not discriminate between first and second hand sources. This suggests that some six- to seven-year-olds, the age of the youngest group of participants, were not showing epistemic awareness regarding who/what would be a better source, depending on what they want to know. In the current study, the younger participants may not yet be ranking perceptual evidence above less salient types of evidence. As a result, their prior beliefs were not overridden by the higher ranking perceptual evidence – observing that weight does not affect distance travelled. It should be noted that the task used by Fitneva (2008) was linguistically based, whereas this study is based on the demonstration of causal

understanding. It is likely that the same mechanisms are used to evaluate source reliability. One would expect this to be the case as it has been observed in young children; four-year-olds have been observed making causal predictions in similar ways in different domains (Schultz & Gopnik, 2004). This does not, however, explain the lack of an age difference with regards to source reliability.

Another potential explanation for why age, contrary to the prediction based on previous research, did not appear to affect participants' predictions regarding the effect of weight, in light of information from differentially reliable sources, may be explained by the fact that the older you get, the harder it is to let go of your prior beliefs (Gopnik et al., 2017). If there was an age-related effect then one would expect more of the older participants to adjust their predictions relating to weight when provided with unexpected evidence from a high reliability source, compared with younger participants, but this was not the case. Older participants were capable of overriding their prior beliefs in the face of perceptual evidence, and almost all of them did so, so it is not an issue with strong prior beliefs per se. It is possible that providing an explanation at baseline reinforced the strength of their prior beliefs. There is evidence to suggest that five-year-olds are more likely to prefer hypotheses that concur with their prior beliefs after providing an explanation, and nine- to 10-year-olds are more likely to ignore relevant evidence following explanation of causal claims (Kuhn & Katz, 2009), and that explaining also leads to over generalisations in young adults (Williams et al., 2013).

It may also be possible that the older participants did not find a high reliability source to be reliable *enough*, to override their prior beliefs. They may not have thought a physics teacher was either a suitably reliable source of information in general or in this specific causal system. There is some evidence to suggest that younger participants perceived the high reliability source as more reliable compared to the older participants; the two youngest age groups provided higher ratings of perceived reliability for the high reliability source compared with the older groups. There were no age differences in ratings for the low reliability source. Secondary school children do not always have good relationships with their teachers (Feldlaufer, Midgley, & Eccles, 1988), particularly children from lower SES environments. This can impact on their learning, particularly with children from

ethnic minorities (Crosnoe, Johnson, & Elder, 2004; Decker, Dona, & Christenson, 2007), who were a substantial proportion of the population in both school C and school D. If the children's evaluation focused on the *teacher* component of the source rather than the 'physics' component, and many secondary school children participating did not have positive relationships with their teachers, then they may have felt that teachers were not reliable *in general*. Teachers in both secondary schools commented to the researcher informally on the difficulties they faced in dealing with behavioural problems as part of their ongoing teacher responsibilities, and a number of conflicts between teachers and students were also observed during testing at the schools. There may not have been an age-related source reliability effect for that reason.

Alternatively, it is possible that participants did not think that the physics teacher was a highly reliable source, regarding the causal system at hand. That is, although they thought teachers are generally reliable, they did not think that a *physics* teacher would have 'expert' knowledge compared with them. This could be because some participants did not recognise that 'physics' knowledge related to the causal system. It is difficult to ascertain how likely this is. A number of participants spontaneously commented on the apparatus, referring to physics, which suggests that at least some participants were aware that a physics teacher should have expert knowledge regarding system. However, this would only be relevant if those participants heard information from a high reliability source (comments outside of data collection were not recorded so it is not known which source reliability groups they were in). In school D, it was widely known that the teachers who taught physics did not have physics degrees (there was a poster on the wall in the science block, identifying the school science teachers, that included information regarding their university education). On being asked why his reliability score for the physics teacher was so low, one participants from school D derided the knowledge of his physics teacher, who had, according to the participants, made an error in his most recent physics class.

An alternative conception is that some of the older participants may have sought to intuit what was going on in the study. They may have recognised that being told this information (in general) was somehow relevant to the outcome of the study, and

adjusted their predictions as a result of that. Landrum et al. (2015) suggested that learning from other people requires integration of reasoning about the informant's psychological properties, *and* the implications of the information at hand. The fact that the participants only received information from *either* a high *or* low reliability source meant that understanding the goal (looking for source reliability related differences) may not have been accessible to the participants. However, they may have wondered if there was a 'trick' that they needed to avoid, and tried to provide predictions that avoided falling for the trick. Recruitment in school C was provided via a psychology teacher and many of the older participants were studying for psychology A-level. They may have been aware that psychology experiments frequently seek to conceal the purpose of the experiment. This type of metacognition may have led participants to provide predictions that were not related to source reliability, or related to it in unexpected ways. There is some evidence that knowing about methods for understanding human behaviour in psychology experiments alters responses in future testing. Silverman, Shulman, and Wiesenthal (1970) found that, when participating in a deception study and then debriefed regarding the deception, undergraduates differed from participants who participated in an experiment without deception and debriefing. In this case, they found that deception increased the tendency for favourable self-presentation, and compliance with demand characteristics. In another study of the effects demand characteristics might have on participants' behaviour, Nichols and Maner (2008) found that when undergraduate participants knew the experimental focus, there was an increased demand characteristic effect. This tendency to confirm the 'hypothesis' was increased for those with positive attitudes to the experiment/experimenter. In a review regarding the costs of deception in psychology experiments, Ortmann and Hertwig (2002) concluded that experience with deception had the potential to alter experimental performance, and also to generate suspicion and second guessing when participating in later experiments. While participants were asked not to discuss the experiment with their participating peers, some of them might have shared information, which may have impacted on their responding. Taylor and Shepperd (1996) found that participants would often communicate with each other when left alone, even when asked not to, and would then not report this to the experimenter. While this may not be an issue for many experiments, where

participants do not have social relationships, this is not the case in schools. Particularly in secondary school, small groups of friends would volunteer to participate, and those that had participated may have discussed their experience with those that did not.

In conclusion, there was, as predicted, an effect of source reliability, suggesting that source reliability is associated with participants' understanding of the causal system. However, there did not appear to be any age-related effects. This could mean that younger participants show epistemic awareness regarding what sources know, and how that relates to the causal system at hand. However, some younger participants did show evidence that they may not have understood the implications of the unexpected information they had heard regarding the causal system, and/or not be able to override their prior beliefs, given they ignored the observational evidence. This failure to incorporate relevant evidence into reasoning regarding a causal system suggests that age-related differences do exist. An alternative view is that older participants were not more likely to discriminate between the differentially reliable sources, because they were unable to override their prior beliefs regarding the system, or because of possible prejudices about teachers being knowledgeable. The high reliability source was not 'reliable' enough, which inhibited any age advantage there might have been.

### 4.1.2 LANGUAGE

Receptive vocabulary measured by the BPVS scale was, contrary to expectation, not associated with the likelihood of making correct predictions regarding weight. This was despite the fact that the measure used is known to correlate positively with other vocabulary tests, and with individual intelligence tests (Robertson & Eisenberg, 1981). There are a number of potential explanations for this: first, that receptive vocabulary, and potentially language ability, do not affect performance in this task; second, that either the task used to measure receptive vocabulary, or receptive vocabulary itself, were not adequate measures of the elements of language ability that would be associated with making correct predictions; and third, that the assessment used did not appropriately measure vocabulary across all the age groups.

With regard to the first issue, it is possible that the simple nature of the task, providing a prediction regarding how far you think the car will travel, based on intuition, is such that superior language ability is not necessary to perform it appropriately. The work of Gopnik and colleagues examining children's understanding of unfamiliar causal systems found that children were capable of making correct predictions regarding causal outcomes, even as young as two years of age (Gopnik et al., 2001). However, an improvement in performance with age was also not found, and it is possible that similar types of explanations could be found for why there was no benefit for language ability.

As to the question of whether receptive vocabulary was an appropriate measure as an indicator of language ability, Bryant et al. (2017) found that both reading comprehension and vocabulary explained a substantial amount of the relationship between SES and science attainment. Vocabulary is also a strong predictor of reading comprehension skills (e.g. Lervåg & Aukrust, 2010). It seems unlikely that the requirement of receiving and understanding verbal information is not impacted by language ability.

This leads to the third issue, that the assessment used did not appropriately measure vocabulary across all the age groups. There was a negative relationship between age and BPVS score, where, as the participants got older, the BPVS scores decreased. This is likely to be an artefact of bias in the school populations, with the secondary schools having more students from a lower SES environment, compared with the primary schools. Students from lower SES environments tend to perform less well in many aspects of attainment and in language or IQ tests (Bryant et al., 2017). Added to this, the secondary schools also have a higher proportion of students for whom English was a second language (school C – 49%, school D – 60%), in comparison with the primary schools (school A – 41%, school B – 36%). It is likely that having English as a second language impacted negatively on the vocabulary test, such that it did not give an appropriate measure of 'language ability'. Mahon and Crutchley (2006) found that monolingual children had higher BPVS scores than those with English as a second language. Melby-Lervåg and Lervåg (2014) conducted a meta-analysis looking at language comprehension in monolingual and English as a second language learners. They found a large deficit in language

comprehension, as well as a medium-size deficit in reading comprehension. The largest group differences in relation to language comprehension were studies with samples from low SES backgrounds, and where the first language was only used at home. Together, this suggests that the secondary school and primary school sample differed on characteristics that could add systematic variance into the data, such that differences, or lack thereof, could be explained by differences in the two samples. This problem could explain the lack of an effect of language ability (as measured by the BPVS scale), as well as the lack of an effect of age.

In conclusion, whilst there appears to be no relationship between language ability and making correct predictions, it is possible that there are systematic differences between the primary school and secondary school samples, which may explain the lack of relationship.

### 4.1.3 Degree of conviction

According to the predictions, the univariate analyses showed that participants who made correct predictions were more likely to provide higher ratings of conviction, compared with participants who made incorrect predictions after hearing unexpected information regarding the causal system. Participants who received information from the high reliability source appeared to be driving this effect. This is in keeping with what has been found in studies involving adults (e.g. Hahn et al., 2009; Hahn et al., 2005), which found that adults rated arguments from higher reliability sources as being more convincing compared with arguments from lower reliability sources. There was also, as expected, an effect of degree of conviction after participants had observed that the unexpected information was 'correct'. In this case, it appeared to be the low reliability source and no information groups who were driving the effect. This may be because participants feel more conviction making a prediction that concurs with what they have observed, than one that does not. The difference was not observed for the high reliability source group, although it was expected that a difference would be evident. Furthermore, of the participants who made incorrect predictions in all three groups combined, 56% were in the youngest age group. As was discussed previously, it was not clear that the younger participants were able to inhibit their prior beliefs, or fully understood the causal system, and this may have influenced their ratings. However, the number of

participants who made incorrect predictions after they had witnessed that weight does not affect distance travelled, was fairly small in all three groups, so these results must be regarded with caution.

In addition, due to the small sample and relatively small number of participants who made correct predictions, it was not possible to do a multivariate analysis including the source reliability by correctness of prediction interaction term. As such, it was not possible to demonstrate that source reliability by correctness was independently associated with making a correct prediction once the other factors have been taken into account.

In conclusion, as expected and similarly to the adult literature, this study found preliminary evidence to suggest that participants show a higher degree of conviction in their correct predictions when they concur with information that comes from a high reliability source, in comparison with participants who make incorrect predictions that disagree with information from the same source. Even though these participants disregard the relevant information, it did appear to have impacted on their beliefs regarding the system, if only to decrease their certainty regarding how the system works.

## 4.2 Explicit understanding of the causal system (weight explanation)

As predicted, source reliability was relevant in participants' explanations regarding familiar causal systems. Participants who received relevant information from a high reliability source were more likely to incorporate that information into their explanation, compared with participants who received no information. Participants who received information from low reliability sources did not do this. This suggested that participants were able to demonstrate epistemic awareness regarding what a source might know, and utilise that information in an explicit way, by providing a correct explanation regarding what has occurred.

After participants had witnessed that the weight of the car had no effect on distance travelled, contrary to what was predicted, there was no advantage to hearing

relevant information (from either source). This is unexpected given that there is evidence to suggest hearing relevant information from an adult (in this case a parent) aids understanding in scientific reasoning, particularly when participants also observed the correctness of the explanation (Philips & Tolmie, 2007). It is possible that the limited information the children received in this study precluded the learning that occurred in Philips and Tolmie (2007), or that the information exchange between the parent and child is privileged in some way.

However, one methodological problem with the weight explanation data was that, when children had just observed that weight did not affect how far the car travelled, they commonly responded to the request for an explanation with "because I just saw it" or words to that effect. This response was quite resistant to change, and led to children having poorer explanation scores after witnessing that weight did not affect how far the car travelled (when one would expect the explanations to be better). This may provide a reason why the number of correct explanations barely increased after observing that weight did not affect distance travelled (9% versus 12%), as well as why no differences were observed between source reliability and no information groups as predicted. It is not possible to know whether participants who provided a "because I just saw it" response were, in principle, capable of generating a correct explanation. However, it seems likely that at least some of the participants would be able to do it, especially the older ones.

This issue clouds understanding of one of the main difference between participants' predictions and explanations regarding the effect of weight. This is because the majority of participants made correct predictions after observing that weight did not affect distance travelled. However, this was not the case for explanations. One might expect differences between performance making predictions, which could demonstrate *implicit* understanding, and performance making explanations, which could demonstrate *explicit* understanding. For example, the participants who received *no* relevant information regarding the effect of weight might easily be able to generate correct *predictions* after observing that weight did not affect distance travelled. However, generating a correct explanation regarding a causal system, one that directly contravenes their prior beliefs, may be more difficult if they have not been provided with information that could structure their explanation. The only

indication that this might be the case was the fact that none of the participants in the 'no information' group provided a correct explanation, although some did provide ambiguous explanations.

In summary, as predicted, there was an effect of source reliability, whereby participants who received information from a high reliability source were more likely to incorporate that information into their explanations regarding the effect of weight. However, there was no advantage to hearing relevant information regarding the effect of weight on explanations, after participants had observed that weight did not affect distance travelled. It is likely that this lack of an advantage is due to a methodological issue, whereby many participants resorted to providing an explanation by reporting that they have just observed that weight did not affect distance travelled. The fact that the large majority of participants did not provide correct explanations was the main difference between performance regarding prediction, and explanation. Although resorting to an easier type of explanation may indicate that participants found generating explanations harder than generating predictions, that conclusion cannot be drawn at this point.

### 4.2.1 Age

Contrary to expectation, age was not relevant to the quality of explanations, since explanations were not better for older participants. This was unexpected given that (in a homogeneous sample) language ability improves with age, and explanation relies on fundamental language skills. However, there was also no effect of age for prediction. In particular, if some older participants were either unable to override their prior beliefs, or did not find the high reliability source reliable enough, then they would not be expected to make correct explanations regarding the effect of weight either. If that was the case, then any advantage of age would not be observed as they would have continued to provide incorrect explanations. Although these 'incorrect' explanations may be better than the 'incorrect' explanations of the younger participants, that was not the question of interest.

Furthermore, as discussed in section 4.1.2, there were more participants with English as a second language in the older age groups, which is known to affect language ability (e.g. Melby-Lervåg and Lervåg, 2014). Given that explanation relies

on fundamental language skills and knowledge of English, it is possible that the larger number of participants with English as a second language contributed to the lack of age-related relationship on correctness of explanations, by negatively impacting on performance in the older participants to a greater extent.

Given that the only age-related differences observed with regards to the prediction data were following observation that weight did not affect distance travelled, it might be expected to see a similar difference here. However, this lack of difference may relate back to the default explanation of many participants, "because I just saw it". This response may have been used by participants because they were not sure how to provide a correct response, given evidence that prior beliefs were not 'true'. If that was the case, then any age-related differences may be masked due to the low correct explanation rate that resulted.

Another issue may be that the provision of an explanation for their original beliefs regarding weight increased the chances of many participants ignoring relevance evidence, resulting in fewer correct explanations (Kuhn & Katz, 2009; Williams et al., 2013). Here, providing an explanation *before* being given the new evidence (at baseline) may have decreased the attention paid to the new information, at least by some participants. Participants in Kuhn and Katz's (2009) study repeatedly provided explanations, directly followed by a prediction task, over a number of time periods. That study might provide evidence for an argument to suggest that the provision of an explanation *after* hearing unexpected information may impact on the explanation *directly following* observing that weight does not affect distance travelled. This is because these events occurred during a single session. This was not the case for participants who provided an explanation in the first session with a delay before the second session. There was usually a delay of at least a few days between provision of the first explanation regarding the effect of weight at baseline and receiving the information which occurred at the beginning of the second session for the primary school participants (and often more, as their availability was determined by the school). However, testing in secondary school was different in that it did not take place during class time, so testing sessions were frequently closer together. This was done to ensure that participants completed both sessions (secondary school participants were required to remember their testing

appointments, so keeping them closer together increased full participation). Given the shorter period between providing an explanation at baseline, it is possible that the effect of explanations increasing prior beliefs may have been stronger in the older participants, and this may have resulted in a lack of age differences.

Another potential issue that may have influenced age-related differences was that the explanations provided may not have reflected the younger participants' true ability to provide explanations. They could merely have reflected the fact that the participants in the source reliability conditions were given the correct explanation "weight does not make a difference to how far the car travels", which they then utilised. As stated earlier, all the participants who made correct explanations were from the high and low reliability groups. That is, they had already heard the explanation when asked to generate an explanation for an outcome that directly contravened their prior beliefs. Furthermore, their explanations frequently involved "it doesn't make a difference…", which directly reflected the language used in providing the relevant information. This supports the idea that the information provided facilitated their explanation. Philips and Tolmie (2007) found that participants provided better explanations with parental support, when learning about a science problem. Fang and Wei (2010) found an improvement in scientific reasoning for participants who received a home science reading program, as well as in school instruction. Whilst it was not possible to identify the direct impact of the home science reading intervention, the authors concluded it also benefitted scientific reasoning (see also Gerber et al., 2001; Leichtman et al., 2017). Some evidence to support the theory that many of the weight explanations involved participants simply repeating the source information is that only older participants provided correct explanations at baseline. Furthermore, the majority of participants who provided higher scoring explanations came from secondary school. Although there were too few correct explanations to do in-depth analyses, this suggests that there may be age-related differences in the quality of explanations provided, as has been found in earlier literature (Lombrozo, 2006).

In summary, there were no age-related differences in explanations. This was unexpected given that generally language ability improves with age. However, similarly to the findings regarding prediction, the small number of participants

providing correct explanations, coupled with a relatively small sample, made it difficult to draw strong conclusions in this case.

### 4.2.2 LANGUAGE

Contrary to expectation, no significant relationship was found between receptive vocabulary and the correctness of explanations. This was the case even though generating explanations relies on fundamental language skills (Norris & Phillips, 2002), in a way that generating predictions does not. Support for this has been provided in previous research, where children are first observed providing predictions regarding causal relations at a younger age, compared with providing explanations. Children as young as two years of age have been observed predicting outcomes regarding unfamiliar causal systems (Gopnik et al., 2001), whereas similar research looking at the role of explanation begins with children aged five years old.

However, as discussed previously (section 4.1.2), there may be an issue with the assessment used to collect information regarding participants' receptive vocabularies in this study. That is, compared to the primary schools, the secondary schools involved in the study had a higher percentage of students who came from lower SES environments, and a higher percentage of students with English as a second language. Both these factors can have a negative effect on participants' scores on measures of receptive vocabulary (Bryant et al., 2017; Melby-Lervåg & Lervåg, 2014). This may then impact on any relationship there might be between language ability and provision of a correct explanation, whereby younger participants had higher receptive vocabulary scores and older participants had lower receptive vocabulary scores on average, resulting in no relationship between receptive vocabulary score and correctness of explanation

Another potential issue is that there were too few correct explanations provided by participants of any age for any relationship to be observed. This led to the development of a binary outcome measure, where participants were deemed to have made either a 'correct' or an 'incorrect' explanation. However, the degree of variation in responses was lost by coding the data in this way. If the relationship between language and explanation performance was small, it may not be observable when using a binary outcome measure. Any small advantage provided by superior

language ability offered to explanation provision is likely to have been lost when all correct responses were grouped together.

In summary, although language ability, as measured by receptive vocabulary, was not related to performance providing correct explanations, it is possible that this is explained by differences in the sample, and/or type of language measure used.

### 4.2.3 Degree of conviction

As predicted, after they had heard relevant information regarding the familiar causal system, participants reporting a higher degree of conviction in their predictions were more likely to have made a correct explanation compared with participants who had a lower degree of conviction. Contrary to what was found with the prediction data, this was not the case after they had observed that weight does not affect distance travelled. However, as discussed previously, this result cannot be regarded as a true representation of the participants' abilities to generate correct explanations, as many gave a 'because I just saw it' type of explanation. As such, many participants included in the incorrect explanation group may have been confident in their prediction, thereby providing a higher degree of conviction ratings. This could result in a smaller difference between degree of conviction for those who provided correct and incorrect explanations.

However, contrary to what was predicted, there was no effect of source reliability on the degree of conviction data, in contrast to what was found with the data on prediction. Participants in the high reliability group did not provide higher ratings of degree of conviction if they provided a correct explanation, compared with if they provided an incorrect explanation. It should be noted that the rating was provided regarding participants' predictions, not their explanations. This corresponds with the literature on source reliability understanding, where participants are usually asked to show preferences rather than provide explanations (e.g. Birch et al., 2008; Koenig et al., 2004; Jaswal & Neely, 2006). However, in contrast to this, studies looking at source reliability understanding in adults frequently asked adults to evaluate the strength of explanation type arguments (e.g. Hahn et al., 2005; Hahn et al., 2009). It may have been that the factors involved in evaluating the accuracy of a prediction may be less complicated than those involved in evaluating the

appropriateness of an explanation. For example, for prediction it is necessary to decide where the car will land, based on prior beliefs, and any new relevant information that has been provided (in this case from differentially reliable sources) that the participant has to decide whether to believe. For the explanation, there are other factors involved, such as language ability (Norris & Phillips, 2002), specifically as it relates to generating explanations. It is possible that some participants were capable of generating correct predictions in concordance with evidence from high reliability sources, but not correct explanations. As such their high degree of conviction ratings would be included in the incorrect explanation group, thereby decreasing the difference between participants who made a correct and incorrect explanation in the high reliability source group. There is only a small difference between number of participants who provided correct predictions versus explanations in the high reliability source group, but it may be enough to remove the effect given the small sample size.

In summary, although participants who provided a correct explanation had higher degree of conviction ratings, contrary to the literature there was no effect of source reliability by degree of conviction on correct explanations. However, given providing correct explanations appears to be more difficult than providing correct predictions, it is possible that degree of conviction ratings does not map directly onto correct versus incorrect explanations.

## 4.3 GENDER

As expected, there was an effect of gender with boys more likely to provide both correct predictions and explanations regarding weight after receiving information from differentially reliable sources. This concurs with the previous literature which finds that males are more likely to perform more highly in science at school (Curran & Kellogg, 2016; Nunes et al., 2017; Quinn & Cooc, 2015). The current study was, however, unbalanced in terms of gender, which may have affected the outcome. There were more females (62%) than males overall, but this difference was complicated by the fact that the gender imbalance differed by age; the primary school sample had a slightly higher proportion of males while the secondary school sample had a substantially higher proportion of females. So it is possible that the

gender differences observed here could be explained by age-related differences in attitudes toward source reliability in males versus females. That is, the oldest participants, more of whom were female, were more likely to disregard information that does not concur with their prior beliefs. However, younger participants were less likely to disregard information that did not concur with their prior beliefs. This could result in the finding that males showed better performance with regards to prediction compared with females overall. Another alternative is that the older participants, more of whom were female, were also more likely to come from lower SES environments, found in previous research to be associated with lower attainment in science at school (Nunes et al., 2017). The fact that the males also had higher average BPVS scores may provide some additional support for this as language ability is related to SES (Nunes et al., 2017). As such it is difficult to conclude that the gender differences observed here represent typical gender differences observed in the literature. More research would be necessary, ensuring a sample of different ages that included the same proportion of males and females at each age.

In summary, although gender differences were observed, is not clear that they are actual gender differences, or have been influenced by systemic differences contained within the sample.

## 4.4 UNDERSTANDING THE CAUSAL SYSTEM

As expected, the majority of the participants made correct predictions regarding the causal variables at baseline, suggesting that they understood how the causal system worked. They appeared to understand the effect of height better, compared with starting point on the incline, and surface friction of the incline. Very few participants made incorrect predictions, in the wrong direction, and there were no age-related differences. Overall this suggests that the causal system was generally well understood.

However, observing that weight did not affect the distance that the car travelled was related to participants' understanding of the causal variables. Although there was little difference between participants' causal variable predictions at baseline and after hearing relevant information regarding the effect of weight, there was a

difference after observing that weight did not affect the distance travelled. In essence, many participants shifted from making correct predictions to incorrect predictions, by predicting that one or more of the causal variables had no effect on distance travelled (similarly to what they had just observed for weight). They were most likely to do this for friction, which makes the most sense, as surface friction has the least impact on distance travelled of all three variables.

Some participants may have held onto their beliefs that weight does, in principle have an effect on distance travelled, but not for this equipment. They may have reasoned that, if it was the case in this instance, then it was possible that it was the case for other variables as well, and adjusted their predictions accordingly. If this was true then one might expect to see an effect of source reliability, whereby hearing unexpected information regarding the causal system from a high reliability source indicated to the participant that what they had observed related only to weight, and not the other variables. The numbers are too small to conduct appropriate analyses in this case. However, this does suggest that revising beliefs can be very difficult when they are counter to prior beliefs. At least some participants are likely to revise their entire causal understanding of a particular system, rather than separating out variables that play a causal role and variables that do not. This suggests that those participants have not yet gained effective strategies for understanding causal relations. It is possible that their control of variable strategies would be poorer than participants who did not readjust all their beliefs regarding the causal system.

In summary, participants seemed to understand the causal system, with height being the best understood variable. However, for many participants, observing counterintuitive evidence regarding one variable in the causal system, impacted on their understanding of the whole causal system, rather than just that variable.

## 4.5 PRACTICE TRIALS

Prior to testing, participants had the opportunity to play with the causal system. Some exploratory analyses were done on the data relating to the choices of trials they made. Younger participants were more likely to repeat trials compared with older participants. This contributes towards the suggestion that younger participants have less well developed executive function skills (Diamond, 2013), and

as a result may be less likely to remember which trials they have already done. They may also want reassurance by repeating a trial rather than trying a different strategy.

Use of extreme set up trials, where participants sought to discover the furthest and least distance a car could travel, could give information regarding how the causal system worked. Many participants tried to gain this type of information, with differences based on both gender and receptive vocabulary score. However, as the males in this sample had higher receptive vocabulary scores on average, it is difficult to speculate which of these variables might have been influencing their use of extreme set up trials. It is possible that males have more experience using toy cars on inclines in play from a young age (Todd, Barry, & Thommessen, 2017), during which finding how far a car can travel is a common pastime, and so they engaged in similar activities here. This is in contrast to females who are likely to have played fewer such games, and therefore may have fewer preconceived ideas regarding 'play'. Alternatively, if the BPVS score is seen as a marker for intelligence (Robertson & Eisenberg, 1981), then one might interpret the difference as engaging in information finding regarding the causal system, where intelligence is a predictor of better scientific reasoning.

Although some participants engaged in spontaneous control of variable type trials, there did not appear to be a relationship between this and age, receptive vocabulary or gender. Cook et al. (2011) found that, while four-year-olds did engage in spontaneous testing of variables, they did so only when faced with ambiguous evidence. When faced with unambiguous evidence, they played indiscriminately. In this case, it is possible that, for some participants, prior beliefs were functioning as unambiguous evidence regarding how the causal system works, and so did not feel the need to test the system. The lack of age differences could reflect the strengthening of prior beliefs over time (Gopnik et al., 2017).

## 4.6 LIMITATIONS

### 4.6.1 SAMPLE

#### 4.6.1.1 INDIVIDUAL DIFFERENCES

Receptive vocabulary was the only measure of individual difference collected in this study. One reason why this particular measure was chosen was it could be delivered relatively quickly compared with other measures of language ability, since constraining the length of time testing took was an issue. However, individual differences such as IQ, reading comprehension, and scientific reasoning ability are known to be related to science attainment; it has been reported that around 40% of the variance in scientific attainment can be explained by IQ, reading comprehension, and scientific reasoning ability (Nunes et al., 2017). Nunes et al. (2017) concluded that the relevance of SES for students' science attainment was largely dependent on differences in reading comprehension and scientific reasoning. They also pointed out that much of the research looking at the role of language ability and success in learning science does not take into account IQ, making it difficult to draw conclusions relating to cause and effect. As such, it is likely that some of the variance between participants who did or did not change their predictions and explanations in light of the new information might be further explained by these factors. Also, as the study did not have access to measures of attainment in science, it was not possible to compare our participants' performance in our task with their science attainment. One would predict that there would be a relationship between the two, and it would be important to look for relationships between the two variables in any future study.

Furthermore, information regarding individual SES was not available so could not be included in the analyses. Given the extensive evidence that there is a relationship between SES and science attainment, it is likely that there would be one here also. Having this information would have strengthened the study since the students came from schools with pupils from a range of SES backgrounds. As such, any future research should seek to provide a measure of individual SES.

#### 4.6.1.2 SCHOOL DIFFERENCES

There were a number of differences between the schools that participated in this study which are important to highlight. This is because these differences may indicate that the results related to age in particular, and possibly also gender, need to be treated with caution. Firstly, there were differences in the student population relating to SES. The primary schools were both popular oversubscribed church schools, although school A (providing around two thirds of the primary school participants) fully prioritised church applicants, whereas school B accepted 40% of children based on distance from the school. Furthermore, School A served a well-off middle-class area of London, whereas school B served a much more SES mixed area of London. This was indicated by the comparatively higher percentage of children on free school meals in school B compared with school A, a common indicator of lower SES. School B also had fewer children reaching the expected standard in English and maths compared with school A, which may not be surprising if SES is associated with cognitive achievement throughout life, where cognitive achievement includes IQ, language and school performance (see section 2.1.1.1and 2.1.1.2; Nunes et al., 2017).

However, even more students from both the secondary schools were likely to have come from lower SES environments compared with the primary schools. Secondary school C is a girl's school (providing around three quarters of the secondary school participants), which accepts 75% of its students on distance and 25% on ability. It serves a highly multicultural area of London, indicated by the large number of students with English as second language. The student population is also mixed in terms of SES, and almost half of the children are entitled to free school dinners (see section 2.1.2.1). School D is a boy's school, with girls in the sixth form and Students are accepted largely on distance. It has also a large number of students with English as a second language, and an even larger number of children entitled to free school dinners (see section 2.1.2.2 Table 2-4).

Given the school catchment areas cover such a broad range of SES backgrounds, with a tendency towards younger children having higher SES and older children lower SES backgrounds, it was difficult to draw strong conclusions relating to the absence of age differences. This is because SES is known to affect performance in science

attainment (Nunes et al., 2017). For example, it is possible that there is a difference between how teachers are regarded as sources in higher and lower SES schools, where one might expect more conflict between students and teachers in a lower SES school. For example, SES is one of the predictors of out-of-school suspension and expulsion at the individual and school level (Skiba, Chung, Trachok, Baker, Sheya, & Hughes, 2014). If this is the case, then age-related differences might be more apparent if the older participants had come from higher SES backgrounds, similar to the majority of the younger participants.

These issues highlight the importance of taking SES into account when investigating aspects of cognition that are known to be influenced by SES. It also indicates that a larger sample of schools, relatively matched in terms of their intake, would have made the research more robust. Much of the academic literature that uses fairly constrained experimental design, such as the research on causal understanding, and scientific reasoning in younger children, avoids this problem by mainly using participants from middle-class or upper middle-class environments. However, this research only tells us how children from higher SES backgrounds, who attend nurseries and schools that have time and space to participate in the studies, perform. As the literature on the impact of SES tells us, it is often not the case for all children.

Other factors that may have had an influence on the way that the participants thought about scientific issues include variables such as the style of science teaching, the general school ethos, and the school environment in terms of behaviour management. There were differences in teaching focus between the two primary schools. School A had a strong focus on science, clearly stated on the school website at the time of testing. In contrast, school B had a strong arts and music focus, seeking to embed arts-based teaching methods throughout their curriculum. This primary focus on the Arts might suggest less emphasis on science (for example, there was no mention of science on their website at the time of testing, where, in contrast, school A included a science curriculum). It is also possible that school B teachers come from an arts or music background, and feasibly may have less exposure to explicit scientific reasoning strategies and scientific concepts. Not all primary school teachers have accurate conceptions of forces and motion (e.g. Kruger, Summers, &

Palacio 1990; Narjaikaew, 2013), which may impede the teaching of these concepts in primary school science class. Furthermore, the early verbal environment children are exposed to at school is known to be related to literacy (Connor, Morrison, & Slominski, 2006). Similarly, teachers' use of sophisticated vocabulary at five to six years old predicted children's literacy performance at nine to 10 years old (Dickinson & Porsche, 2011).

Although this research did not focus explicitly on scientific literacy, is possible that sophisticated use of scientific thinking concepts from an early age impact on the development of children's scientific reasoning and understanding. Furthermore, informal learning environments (both in school and out) have been shown to be related to scientific reasoning (Gerber et al., 2001). A school that has a strong arts and music focus may be less inclined to generate scientific language or provide extra informal learning environments known to benefit children's scientific reasoning skills, when compared with a school that has a strong science focus.

Any future research would need to either seek to minimise school differences in relation to the question at hand, or include school as a potential explanatory variable by including a number of schools with different approaches to the teaching of science.

### 4.6.2 Methodological issues

#### 4.6.2.1 The school environment

One methodological issue was the variability in the space that was allocated for the research. Both primary schools and secondary school D had a problem with finding a space to set up the relatively large apparatus (school C had recently had new premises built, so space was not an issue for them). Secondly, schools differed in the extent to which they promoted the study with families. Communication with parents via the children, to obtain informed consent was challenging, particularly in school B and D (who promoted the study less). The consequence of this was that it limited the number of participants, which meant that strong conclusions regarding any statistical effects were less likely to be identified. The small number of schools who agreed to take part precluded using school as a potential explanatory variable.

Although some of the schools had students from lower SES families, it was not possible to determine whether any particular child participating in the research had a lower or higher SES family background. Higher SES is known to predict greater parental involvement in school activities (Hoover–Dempsey, Bassler, & Brissie, 1987), so it is possible that a greater proportion of children from higher SES families were volunteered by their parents. This was particularly likely in primary schools, where participation was potentially more parent-driven, where students were given information on the study to take home to their parents. The decision to participate would be made once the parent had received the information, and the parent could have played a role in their child's decision to participate. In secondary schools, participation was more student driven. Teachers introduced the study to students, who then volunteered to participate (the younger two age groups in secondary school then had to get parents' consent). This process meant that parents did not play such a strong a role in their child deciding to participate in secondary school, where that decision was student-driven (it was possible that the parent could dissuade their child from participating, however). Nevertheless, is was not possible to determine whether SES affected participation in secondary school. Parental involvement has been found to predict children's participation in extra-curricular activities (Anderson, Funk, Elliott, & Hull Smith, 2003), and SES has been related to parental involvement in school activities (Hoover–Dempsey et al., 1987). So SES may also have had an effect in the nature of the sample in the current study. These issues highlight the importance of being able to collect information about individual level SES in any future research.

#### 4.6.2.2 Test question phrasing

Another methodological issue concerns the phrasing of the question regarding participants' explanation for what they have observed. During this study, participants were asked "why do you think that?" The goal was to get them to provide an explanation for what they believed regarding the effect of weight. This resulted in participants mostly providing explanations as expected. However, some participants, after having observed weight did not affect distance travelled, provided only a perceptually based explanation. This is not incorrect as a response to the question "why do you think that?" However, it does not reveal their beliefs

regarding the effect of weight. One potential solution to this problem would be to reframe the question, in order that it focuses participants' attention more closely on the causal relations in the system. For example, instead of asking participants "why do you think that?", with regards to their different predictions regarding weight in the causal variables, one could ask them "what do you think the effect of weight is?" This is more likely to lead them to frame their reasons based on the implications of the evidence they have just observed, rather than the evidence they had just observed.

#### 4.6.2.3 Correct versus ambiguous explanations

A second and related methodological issue was that, for the purposes of conducting analyses and given the small number who provided correct explanations, they were combined with responses from participants who provided ambiguous explanations (see section 2.6.4). Whilst there was some evidence to suggest that ambiguous explanations were attempts to provide correct explanations regarding the effect of weight, it was not an ideal solution. Any future research would need to make sure that the sample was large enough for appropriate analyses, even in the case that relatively few participants provided correct explanations.

#### 4.6.2.4 Source reliability manipulation

Another potential issue was whether the source reliability manipulation was pertinent enough. A lack of pertinence may have been why only around a third of participants in the high reliability group paid attention the new information regarding the effect of weight, as well as minimal age-related effects. It could be that if the expertise of the source had been made more pertinent, then a stronger effect of source reliability might have been identified, particularly with older participants. The child research literature using the selective trust paradigm, and the adult research literature using the argument strength paradigm, both make the source reliability manipulation fairly obvious. In selective trust tasks, participants were required to gather information on the reliability of the sources during testing, before being asked to make judgements based on information provided by the sources (e.g. Koenig & Harris, 2005). This gave them some time to reflect on the reliability of the sources before being asked to participate in the test phase of the experiment. Furthermore, adults were often provided with booklets containing a number of

arguments which they are expected to rate, where source reliability, for example, is manipulated. There was usually not a time limit. However, in the current study participants were given information from differentially reliable sources and then almost immediately asked for a prediction and explanation regarding weight. It is possible that some participants might have needed time to incorporate the new information, adjust their beliefs, and generate new predictions or explanations. Howe et al. (1992) found that learning and integrating counterintuitive information learned in science benefits from discussion and can continue long after the learning event has concluded (in eight- to 12-year-olds).

Another potential issue was that it was not always clear how much attention the child was paying at the time the information was delivered by the source. The experimenter sought to make sure the child was paying attention, asking them what they thought about the information afterwards, but many children responded to that question with "I don't know". This made the degree to which they had processed the information unclear. "I don't know" could indicate confusion regarding the new information as it relates to the causal system, or indicating that they have not engaged with the information at all.

One possible way forward would be to generate a visible manipulation to accompany the verbal one. Presenting information both visually and orally is known to support learning (Mayer & Anderson, 1992), and may function to make source reliability manipulation more pertinent, and attention-grabbing. One could for example have a video of a teacher in a lab coat, and a young child, both generating the relevant information (in a classroom context). If the failure to adjust predictions and explanations to reflect source reliability was related to the source, not being pertinent enough, then seeing such a video may emphasise the difference between sources much more.

It was not possible to assess whether or how source reliability interacts with observational evidence regarding understanding of causal relations. This was because there were so few incorrect predictions following observation that weight did not affect distance travelled (at the third time point), that any differences between sources would not be observable. Addressing this question would require

a more complex design, making it harder to implement (and children to understand). In the future, one could manipulate the frequency with which the physical evidence concurred (e.g. always, often, sometimes, never) with the verbal information provided by the source. The implications of observational evidence for causal understanding is not always clear, and verbal information can facilitate learning (Philips & Tolmie, 2007). Given this, one might expect to find that source reliability supplements observational evidence, such that a difference is observed with intermediate frequencies (often, sometimes, never), where more errors are likely to be made.

## 4.7 Future Directions

In order to eliminate some of the limitations outlined in the previous section, future research would be more robust if conducted in direct collaboration with schools.

The experimental procedures could be designed in collaboration with teachers so that the experimental paradigm would be shifted into the classroom. This would entail working directly with the school, and teachers, to investigate the impact of source reliability in classroom environments. As there was an interest in developing an understanding of source reliability to inform teaching and learning, the experimental environment should be as close as possible to the environment of interest.

For example, in primary schools, students might be asked to work out the relevant causal variables in the causal system used in the current study (motion on an incline). They would be able to play with the system, and then generate a baseline set of predictions, by filling in a typical class worksheet. Children might then see one of two short videos that explained how weight does not affect how far the car travels, one by a younger peer, and the other a teacher (both from school). Children would then fill in further worksheet stating what they predict, and an explanation and, after that, they would be asked to assess the veracity of the statement made by either the teacher or the peer. Some guidance would be given as to how to do a fair test, which they would then be asked to do on weight, followed by further prediction and explanation. Such a design more closely mimics activities that normally occur in the

classroom, so would give us a better idea of how children reason about source reliability in situ.

In secondary schools, one could take this further, as both teachers and students could be used as sources. A physics teacher, a non-physics teacher, student peers of the same age or younger/older could all be used. Students can potentially learn from many different sources in the school environment and it would be useful to know how students utilise information from these sources.

Other variables that could be investigated include doing the procedure singularly, or in groups, which are known to benefit learning (Howe et al., 1992). One could examine the impact of cognitive abilities and academic attainment, the impact of teaching methods, and different schools on the relevance of source reliability. Doing this in direct collaboration with schools would help to solve the number of problems. Schools would hopefully encourage participation, which would increase the sample size per school. Designing the experimental procedure with teachers will increase the validity of the design, in that the procedure would be closer to a typical lesson. This is important, as if one wants to know what students do in their real-life environments, then the experimental procedure should mimic that as much as possible.

With the collaboration of schools, it may be more likely to be able to collect individual demographic information, along with academic achievement and possibly other measures of individual difference (as long as parental consent is given). Ideally, one would want to use demographic information that was readily available to the school, though data protection issues would still need to be addressed. This could decrease the amount of time devoted to each student, which was an issue, especially in secondary school where teachers do not like students to miss class time. Furthermore, if one identified subcategories of students, who appeared not to be paying attention to appropriate sources, then one could implement interventions to counteract that. However, one would want to be able to identify these students using currently available demographic information, rather than using specially designed cognitive ability tests which are usually not used or available in schools. With school collaboration, it is likely to be easier to repeatedly

assess performance over multiple sessions, to assess changes in strategy and understanding. Given peer learning frequently occurs after time has elapsed (Howe et al., 1992; Howe, McWilliam, & Cross, 2010), assessment after time has passed would also be very important.

Another crucial area for further investigation is to incorporate factors such as SES into research looking at the development of scientific reasoning. Given that the SES attainment gap in science is so persistent, understanding how to boost the attainment of poor performing students should be of the utmost importance. It is possible that participants from lower SES environments interact with source reliability differently from those in higher SES environments. For example, levels of expert commentary in news stories reinforce SES-based differences in political knowledge in adults, where a knowledge gap has been observed in this arena (Jerit, 2009), similar to that observed in scientific attainment. Jerit (2009) found evidence that suggested the manner in which the media cover political issues could affect disparities in knowledge across SES groups. Disparity appeared to increase when expert commentators were employed to cover political issues. In contrast to this, presenting more contextual information, such as providing more information on the historical, social or political context of important events, decreased the political knowledge gap. Jerit (2009) argued that the disparity could be related to differences in educational opportunity. For example, those who grew up in a higher SES background, would be more likely to have had access to better education, and may find it easier to understand news articles with abstract concepts, technical subjects, and infrequently used words. Feelings of alienation towards reputed experts may pervade source reliability evaluation in lower SES adults' everyday lives, and affect their judgements regarding experts and other high reliability sources, in general. Furthermore, if Jerit (2009) correctly ascribes the disparity to differences in educational opportunity, then these beliefs are likely to begin during the years of education, and may in part contribute towards the science attainment gap.

One goal of science education should be to produce good scientific reasoners. Ideally, school leavers should have an evaluativist epistemological understanding, which would enable them to engage in educated decision-making. However, a recent report from the Confederation of British Industry (2015), suggested that science

education is lacking in the UK, especially in primary school. The survey suggested that many primary school teachers believe that science is becoming less important, and give it less time in the curriculum. This is problematic, as a decrease in interest in teaching science effectively could lead to poorer scientific thinking-skills going into secondary school. There is a big jump in cognitive and academic language proficiency (separate from interpersonal communication skills) that is required to function effectively in secondary schools (Cummins, 1980). Less attention paid to science in primary school would increase the gap, particularly for children from lower SES environments, and/or homes where English is a second language.

Furthermore, another factor that may affect the development of scientific thinking, is the methods of assessment. Currently in the UK, based on annual nationally standardised tests at specific ages, both primary and secondary schools are ranked, and this information is made available to the general public. This has led to the practice of "teaching to the test" where teachers teach their students how to pass the test. As pointed out by Stanovich (2011), this is particularly problematic when it raises test performance without affecting the underlying construct being assessed. This is in contrast to teaching to the test where the underlying skills taught to improve test scores also improve the ability under question. Consider reading - if teachers taught the reading skills that were measured in the 'test', but these reading skills also underpinned reading ability, then teaching to the test in this instance would also be improving reading ability. In the same vein, if in physics children are taught the appropriate answers to particular questions (children are often given homework that comes from GCSE exam papers), but knowing the answers to these questions does not underpin scientific reasoning skills, then many children might leave school with less than adequate scientific reasoning skills.

It is important to gain a good understanding of the development of the required scientific reasoning skills and how they develop in the classroom in order that assessments are designed to measure the skills that students take with them when they leave school. Understanding source reliability is one of those crucial scientific reasoning skills.

## 4.8 CONCLUSION

Even though individuals generally have an excellent intuitive understanding of the physical world around them, this does not always transfer itself to explicit knowledge. It is hoped that good science education will provide the foundation for good scientific reasoning skills, that should facilitate decision-making in the everyday world. One of the very important skills in scientific reasoning is the evaluation of source reliability. Both children and adults have been shown to pay attention to source reliability and even very young children have been shown to have epistemic awareness regarding what sources might know.

The aims of this study were to establish at what age children begin discriminate between differentially reliable sources in more naturalistic environments, showing epistemic awareness; to enable a more direct comparison between the adult and child literature on source reliability; to make a direct comparison between the implicit and explicit understanding of a specific causal systems; and to examine the role that language ability and gender might play. To do that, participants from primary and secondary school (aged six to 17 years) were asked to provide information regarding their beliefs about a specific causal system, before and after being given new information from differentially reliable sources, and after carrying out an intervention, whereby they observed the truth of the new information.

Participants did pay attention to source reliability in that participants in the high reliability source group were more likely to make a correct prediction and explanation regarding the causal system. However, there was not an effect of age, although the younger children may have struggled to incorporate the new information. It is possible that this relates to the older participants struggling to inhibit their prior beliefs. However, it may also indicate that the source reliability manipulation was less effective with older participants. SES factors, which could not be taken into account in the analyses, may also have been at play.

There was a relationship between degree of conviction, and source reliability; participants providing a correct response in the high source reliability group were more likely to report a higher conviction in their response compared with incorrect participants. This allowed comparison with the adult literature on source reliability,

which frequently utilises ratings for argument strength to assess participants' understanding of source reliability, and find differences in argument ratings, dependant on source reliability.

Whilst the lack of age-related differences made it difficult to compare the developmental trajectory of implicit and explicit causal understanding, there was one major difference that could be observed. Participants were much more capable of providing correct predictions, drawing on their implicit causal understanding, than they were of providing correct explanations, drawing on their explicit causal understanding.

Finally, gender may have played a role in performance. However, there are confounding factors that made it difficult to draw conclusions in this study, which may also explain the lack of a relationship between language ability and performance. SES is likely to have influenced performance as well, but was not examined here.

Both children and adults are faced with the never-ending stream of information in the 21st-century that is unprecedented in this world of 'fake news'. It is crucially important in this day and age to be able to evaluate sources, and incorporate the information into reasoning about the world dependent on the reliability of the sources.

# REFERENCES


Ahmed, S. F., Tang, S., Waters, N. E., & Davis-Kean, P. (2018). Executive function and academic achievement: Longitudinal relations from early childhood to adolescence. *Journal of Educational Psychology*, *111*(3), 446-458.

Anderson, H., & Hepburn, B. (2016). In *The Stanford Encyclopaedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/scientific-method

Anderson, J. C., Funk, J. B., Elliott, R., & Smith, P. H. (2003). Parental support and pressure and children's extracurricular activities: Relationships with amount of involvement and affective experience of participation. *Journal of Applied Developmental Psychology*, *24*(2), 241-257.

Ardila, A., Rosselli, M., Matute, E., & Inozemtseva, O. (2011). Gender differences in cognitive development. *Developmental Psychology*, *47*(4), 984-990.

Bernard, S., Proust, J., & Clément, F. (2015). Four-to six-year-old children's sensitivity to reliability versus consensus in the endorsement of object labels. *Child development*, *86*(4), 1112-1124.

Berthelsen, D., Hayes, N., White, S. L., & Williams, K. E. (2017). Executive function in adolescence: Associations with child and family risk factors and self-regulation in early childhood. *Frontiers in Psychology*, *8*, 903.

Berthold, K., Röder, H., Knörzer, D., Kessler, W., & Renkl, A. (2011). The double-edged effects of explanation prompts. *Computers in Human Behavior, 27*(1), 69-75.

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences*, *21*(4), 327-336.

Birch, S. A., Vauthier, S. A., & Bloom, P. (2008). Three-and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, *107*(3), 1018-1034.

Blair, C., & Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. *Development and Psychopathology*, *20*(3), 899-911.

Bornstein, M. H., Hahn, C. S., & Haynes, O. M. (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First Language*, *24*(3), 267-304.

Bowerman, B. L., & O'Connell, R. T. (1990). *Linear statistical models: An applied approach*. Belmont, CA: Duxbury.

Braasch, J. L., Lawless, K. A., Goldman, S. R., Manning, F. H., Gomez, K. W., & Macleod, S. M. (2009). Evaluating search results: An empirical analysis of middle school students' use of source attributes to select useful sources. *Journal of Educational Computing Research*, *41*(1), 63-82.

Bråten, I., Ferguson, L. E., Strømsø, H. I., & Anmarkrud, Ø. (2013). Justification beliefs and multiple-documents comprehension. *European Journal of Psychology of Education*, *28*(3), 879-902.

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science*, *24*(4), 573-604.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick*. Cambridge, MA: Harvard University Press.

Bryant, P., Nunes, T., Hillier, J., Gilroy, C., & Barros, R. (2015). The importance of being able to deal with variables in learning science. *International Journal of Science and Mathematics Education*, *13*(1), 145-163.

Bullock, M. (1984). Preschool children's understanding of causal connections. *British Journal of Developmental Psychology, 2*(2), 139-148.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, *63*(4), 515-533.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367.

Clark, C. A., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental Psychology*, *46*(5), 1176-1191.

Confederation of British Industry. (2015*). Tomorrow's World: Inspiring primary scientists.* Retrieved from https://www.stem.org.uk/resources/elibrary/resource/35987/tomorrows-world-inspiring-primary-scientists

Connor, C. M., Morrison, F. J., & Slominski, L. (2006). Preschool instruction and children's emergent literacy growth. *Journal of Educational Psychology, 98*(4), 665-689.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341-349.

Corner, A., & Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, *15*(3), 199-212.

Corriveau, K. H., Fusaro, M., & Harris, P. L. (2009). Going with the flow: Preschoolers prefer non-dissenters as informants. *Psychological Science*, *20*(3), 372-377.

Crosnoe, R., Johnson, M. K., & Elder Jr, G. H. (2004). Intergenerational bonding in school: The behavioral and contextual correlates of student-teacher relationships. *Sociology of Education*, *77*(1), 60-81.

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly*, 14(2), 175-187.

Curran, F. C., & Kellogg, A. T. (2016). Understanding science achievement gaps by race/ethnicity and gender in kindergarten and first grade. *Educational Researcher*, *45*(5), 273-282.

Dale, E., & Reichert, D. (1957). *Bibliography of vocabulary studies*. Columbus, OH: Ohio State University Bureau of Educational Research.

Danks, D. (2007). Causal learning from observations and manipulations. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 359-388). New York, NY: Lawrence Erlbaum Associates.

Dean Jr, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, *91*(3), 384-397.

Decker, D. M., Dona, D. P., & Christenson, S. L. (2007). Behaviorally at-risk African American students: The importance of student–teacher relationships for student outcomes. *Journal of School Psychology*, *45*(1), 83-109.

Department for Education (2013). Science programmes of study: key stage 3. National curriculum of England. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/335174/SECONDARY_national_curriculum_-_Science_220714.pdf

Department for Education (2014, 2 December). National curriculum in England: framework for key stages 1 to 4. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-framework-for-key-stages-1-to-4

Department for Education (2015-16 Cohort). Find and compare schools in England. Retrieved from https://www.gov.uk/school-performance-tables

Department for Education (2015, 6 May). National curriculum in England: science programmes of study. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135-168.

Dickinson, D. K., & Porche, M. V. (2011). Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development*, 82*(3),* 870-886.

Duch, H., Fisher, E. M., Ensari, I., & Harrington, A. (2013). Screen time use in children under 3 years old: A systematic review of correlates. *International Journal of Behavioral Nutrition and Physical Activity*, *10*(1), 102.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test-revised*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, D. M. (2009). *The British picture vocabulary scale*. Brentford, UK: GL Assessment Limited.

Dunn, L. E. V. I., Dunn, L., Whetton, C., & Pintillie, D. (1997). *British picture vocabulary scale: Revised.* Windsor, UK: NFER-Nelson.

Durkin, K., & Shafto, P. (2016). Epistemic trust and education: effects of informant reliability on student learning of decimal concepts. *Child Development*, *87*(1), 154-164.

Elliot, C. D. (1982). *The British ability scales. manual 2: Technical and statistical information.* Windsor: NFER-Nelson.

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education*, *28*(5), 491-520.

Fang, Z., & Wei, Y. (2010). Improving middle school students' science literacy through reading infusion. *The Journal of Educational Research*, *103*(4), 262-273.

Fay, A. L., & Klahr, D. (1996). Knowing about guessing and guessing about knowing: Preschoolers' understanding of indeterminacy. *Child Development*, *67*(2), 689-716.

Feldlaufer, H., Midgley, C., & Eccles, J. S. (1988). Student, teacher, and observer perceptions of the classroom environment before and after the transition to junior high school. *The Journal of Early Adolescence*, *8*(2), 133-156.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234-248.

Fernbach, P. M., Linson-Gentry, P., & Sloman, S. A. (2007, January). Causal beliefs influence the perception of temporal order. In D. S. McNamara & J. G. Trafton (Eds.). *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Ferretti, R. P., Butterfield, E. C., Cahn, A., & Kerkman, D. (1985). The classification of children's knowledge: Development on the balance-scale and inclined-plane tasks. *Journal of Experimental Child Psychology*, *39*(1), 131-160.

Fitneva, S. A. (2001). Epistemic marking and reliability judgments: Evidence from Bulgarian. *Journal of Pragmatics*, *33*(3), 401-420.

Fitneva, S. A. (2008). The role of evidentiality in Bulgarian children's reliability judgments. *Journal of Child Language*, *35*(4), 845-868.

Fitneva, S. A. (2010). Children's representation of child and adult knowledge. *Journal of Cognition and Development*, *11*(4), 458-484.

Freier, L., Cooper, R. P., & Mareschal, D. (2017). Preschool children's control of action outcomes. *Developmental Science*, *20*(2), e12354.

Fusaro, M., & Harris, P. L. (2008). Children assess informant reliability using bystanders' non-verbal cues. *Developmental Science*, *11*(5), 771-777.

Galsworthy, M. J., Dionne, G., Dale, P. S., & Plomin, R. (2000). Sex differences in early verbal and non-verbal cognitive development. *Developmental Science*, *3*(2), 206-215.

Gerber, B. L., Cavallo, A. M., & Marek, E. A. (2001). Relationships among informal learning environments, teaching procedures and scientific reasoning ability. *International Journal of Science Education*, *23*(5), 535-549.

Gerstenberg, T. and J. B. Tenenbaum (2017). "Intuitive Theories". In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). New York, NY: Oxford University Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, *111*(1), 3-32.

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories* (Vol. 1). Cambridge, MA: Mit Press.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892-7899.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*(5), 620-629.

Gordon, R., Franklin, N., & Beck, J. (2005). Wishful thinking and source monitoring. *Memory & Cognition*, *33*(3), 418-429.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334-384.

Hackman, D. A., Gallop, R., Evans, G. W., & Farah, M. J. (2015). Socioeconomic status and executive function: Developmental trajectories and mediation. *Developmental Science*, *18*(5), 686-702.

Hagmayer, Y., & Waldmann, M. R. (2000). Simulating causal models: The way to structural sensitivity. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 214–219). Mahwah, NJ: Lawrence Erlbaum.

Hahn, U., Harris, A. J., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*, 337-367.

Hahn, U., Oaksford, M., & Bayindir, H. (2005). How convinced should we be by negative evidence? In L. Barsalou, and M. Bucciarelli (Eds.) *Proceedings of the annual Conference of the Cognitive Science Society*, 27 (pp. 887-892). Mahwah, N.J.: Lawrence Erlbaum.

Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development*, *77*(3), 505-524.

Hast, M., & Howe, C. (2012). Understanding the beliefs informing children's commonsense theories of motion: the role of everyday object variables in dynamic event predictions. *Research in Science & Technological Education*, *30*(1), 3-15.

Hast, M., & Howe, C. (2013). The development of children's understanding of speed change: A contributing factor towards commonsense theories of motion. *Journal of Science Education and Technology*, *22*(3), 337-350.

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology*, *37*(5), 668-683.

Hobbs, G., & Vignoles, A. (2010). Is children's free school meal 'eligibility' a good proxy for family income? *British Educational Research Journal*, *36*(4), 673-690.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368-1378.

Hofmann, W., Schmeichel, B. J., & Baddeley, A. D. (2012). Executive functions and self-regulation. *Trends in Cognitive Sciences, 16*(3), 174-180.

Hoover-Dempsey, K. V., Bassler, O. C., & Brissie, J. S. (1987). Parent involvement: Contributions of teacher efficacy, school socioeconomic status, and other school characteristics. *American Educational Research Journal*, *24*(3), 417-435.

Howe, C., McWilliam, D., & Cross, G. (2005). Chance favours only the prepared mind: Incubation and the delayed effects of peer collaboration. *British Journal of Psychology*, *96*(1), 67-93.

Howe, C., Tolmie, A., & Rodgers, C. (1992). The acquisition of conceptual knowledge in science by primary school children: Group interaction and the understanding of motion down an incline. *British Journal of Developmental Psychology*, *10*(2), 113-130.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2), 236-248.

Inhelder, B. & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York, NY: Basic Books.

Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best preschoolers use past reliability over age when learning new words. *Psychological Science*, *17*(9), 757-758.

Jerit, J. (2009). How predictive appeals affect policy opinions. *American Journal of Political Science*, *53*(2), 411-426.

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, *23*(4), 495-518.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2013*).* Motivated numeracy and enlightened self-government. *Behavioural Public Policy, 1(1), 54-86.*

Kigel, R. M., McElvany, N., & Becker, M. (2015). Effects of immigrant background on text comprehension, vocabulary, and reading motivation: A longitudinal study. *Learning and Instruction*, *35*, 73-84.

Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, *125*(5), 524-543.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*(1), 111-146.

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, *15*(10), 694-698.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*, 1261-1277.

Koenig, M. A., & Harris, P. (2007). The Basis of Epistemic Trust: Reliable Testimony or Reliable Sources. *Episteme*, *4,* 264-284.

Koenig, M. A., & Jaswal, V. K. (2011). Characterizing children's expectations about expertise and incompetence: Halo or pitchfork effects? *Child Development*, *82*(5), 1634-1647.

Koenig, M. A., & Woodward, A. L. (2010). Sensitivity of 24-month-olds to the prior inaccuracy of the source: possible mechanisms. *Developmental Psychology*, *46*(4), 815-826.

Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology*, *64*(3), 141-152.

Kruger, C., Summers, M., & Palacio, D. (1990). A survey of primary school teachers' conceptions of force and motion. *Educational Research*, *32*(2), 83-95.

Kuhn, D. (2001). How do people know? *Psychological Science*, *12*(1), 1-8.

Kuhn, D. (2005). *Education for thinking*. Cambridge, MA: Harvard University Press.

Kuhn, D. (2007a). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education*, *91*(5), 710-726.

Kuhn, D. (2007b). Jumping to conclusions. *Scientific American Mind*, *18*(1), 44-51.

Kuhn, D. (2010). What is scientific thinking and how does it develop? In U. Goswami, *The Wiley-Blackwell handbook of childhood cognitive development (2nd Ed*), (pp. 497-523). Hoboken, NJ: Wiley-Blackwell.

Kuhn, D., & Dean Jr, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, *16*(11), 866-870.

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology*, *103*(3), 386-394.

Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development*, *1*(1), 113-129.

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development*, *15*(3), 309-328.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480-498.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology*, *43*(1), 186.

Kushnir, T., Vredenburgh, C., & Schneider, L. A. (2013). "Who can help me fix this toy?" The distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental Psychology*, *49*(3), 446.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, And Cognition*, *32*(3), 451-460.

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109-111.

Landrum, A. R., Mills, C. M., & Johnston, A. M. (2013). When do children trust the expert? Benevolence information influences children's trust more than expertise. *Developmental Science*, *16*(4), 622-638.

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*, *83*(1), 173-185.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198-212.

Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology*, *21*(1), 59-80.

Leichtman, M. D., Camilleri, K. A., Pillemer, D. B., Amato-Wierda, C. C., Hogan, J. E., & Dongo, M. D. (2017). Talking after school: Parents' conversational styles and children's memory for a science lesson. *Journal of Experimental Child Psychology*, *156*, 1-15.

Lervåg, A., & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry*, *51*(5), 612-620.

Li, J., & Klahr, D. (2006). The psychology of scientific thinking: Implications for science teaching and learning. In J. Rhoton & P. Shane (Eds.) *Teaching science in the 21st century* (National Science Teachers Association and National Science Education Leadership Association). Arlington, VA: NSTA Press.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464-470.

Lucas, A. J., & Lewis, C. (2010). Should we trust experiments on trust? *Human Development*, *53*(4), 167-172.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284-299.

Lutz, D. J. and Keil, F. C. (2002), Early understanding of the division of cognitive labour. *Child Development*, *73*, 1073–1084.

Ma, X. (2008). Within-school gender gaps in reading, mathematics, and science literacy. *Comparative Education Review*, *52*(3), 437-460.

Macris, D. M., & Sobel, D. M. (2017). The role of evidence diversity and explanation in 4-and 5-year-olds' resolution of counterevidence. *Journal of Cognition and Development*, 18(3), 1-17.

Mahon, M., & Crutchley, A. (2006). Performance of typically-developing school-age children with English as an additional language on the British picture vocabulary scales II. *Child Language Teaching and Therapy*, *22*(3), 333-351.

Masson, S., Potvin, P., Riopel, M., & Brault, L-M. (2014). Differences in brain activation between novices and experts in science during a task involving a common misconception in electricity. *Mind, Brain, and Education*, *8*, 44–55.

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology*, *101*(3), 689-704.

Mayer, R. E., & Anderson, R. B. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology*, *84*(4), 444-452.

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction*, *29*, 43-55.

Melby-Lervåg, M., & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first-and second-language learners. *Psychological Bulletin*, *140*(2), 409.

Mildenhall, P. T., & Williams, J. S. (2001). Instability in students' use of intuitive and Newtonian models to predict motion: the critical effect of the parameters involved. *International Journal of Science Education*, *23*(6), 643-660.

Mills, C. M., Legare, C. H., Grant, M. G., & Landrum, A. R. (2011). Determining who to question, what to ask, and how much information to ask for: The development of inquiry in young children. *Journal of Experimental Child Psychology*, *110*(4), 539-560.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49-100.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, M. W., & Caspi, A. (2011). A gradient of childhood self-

control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, *108*(7), 2693-2698.

Morris, B. J., Croker, S., Masnick, A., and Zimmerman, C. (2012). The emergence of scientific reasoning. In H. Kloos, B. J.Morris, and J. Amaral (Eds.). *Trends in cognitive development* (61-82). Rijeka, Croatia: InTech.

Narjaikaew, P. (2013). Alternative conceptions of primary school teachers of science about force and motion. *Procedia-Social and Behavioral Sciences*, *88*, 250-257.

National curriculum assessments: information for parents. (2017, 28 March). Retrieved from https://www.gov.uk/government/collections/national-curriculum-assessments-information-for-parents.

Ng-Knight, T., & Schoon, I. (2017). Can locus of control compensate for socioeconomic adversity in the transition from school to work? *Journal of Youth and Adolescence*, *46*(10), 2114-2128.

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, *135*(2), 151-166.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224-240.

Nunes, T., Bryant, P., Strand, S., Hillier, J., Barros, R., & Miller-Friedmann, J. (2017). *Review of SES and science learning in formal educational settings. A report prepared for the EFF and the Royal Society*. London, UK: Education Endowment Foundation. Retrieved on https://royalsociety.org/-/media/policy/topics/education-skills/education-research/evidence-review-eef-royalsociety-22-09-2017.pdf

Nurmsoo, E., & Robinson, E. J. (2009a). Children's trust in previously inaccurate informants who were well or poorly informed: When past errors can be excused. *Child Development*, *80*(1), 23-27.

Nurmsoo, E., & Robinson, E. J. (2009b). Identifying unreliable informants: Do children excuse past inaccuracy? *Developmental Science*, *12*(1), 41-47.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. New York, NY: Oxford University Press.

Office of the Administration for Children & Families, Office of Head Start (n.d.). *Head start programs.* Retrieved from https://www.acf.hhs.gov/ohs/about/head-start

Ortmann, A., & Hertwig, R. (2002). The costs of deception: Evidence from psychology. *Experimental Economics*, *5*(2), 111-131.

Osman, M., & Shanks, D. R. (2005). Individual differences in causal learning and decision making. *Acta Psychologica*, *120*(1), 93-112.

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216-1226.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. New York, NY: Cambridge University Press.

Philips, S., & Tolmie, A. (2007). Children's performance on and understanding of the balance scale problem: The effects of parental support. *Infant and Child Development*, *16*(1), 95-117.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. *Psychological Review*, *111*(3), 781-799.

Quinn, D. M., & Cooc, N. (2015). Science achievement gaps by gender and race/ethnicity in elementary and middle school: Trends and predictors. *Educational Researcher*, *44*(6), 336-346.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401-11405.

Robertson, J. R., & Eisenberg, J. L. (1981). *Technical supplement to the Peabody picture vocabulary test-revised.* Circle Pines, MN, USA: American Guidance Service.

Ruffman, T., Perner, J., Olson, D. R., & Doherty, M. (1993). Reflecting on scientific thinking: Children's understanding of the hypothesis-evidence relation. *Child Development*, *64*(6), 1617-1636.

Saçkes, M., Trundle, K. C., Bell, R. L., & O'Connell, A. A. (2011). The influence of early science experience in kindergarten on children's immediate and later science achievement: Evidence from the early childhood longitudinal study. *Journal of Research in Science Teaching*, *48*(2), 217-235.

Salmon, W. (1990). Rationality and objectivity in science or Tom Kuhn meets Tom Bayes. *Scientific Theories*, *14*, 175-204.

Salmon, M. H., Earman, J., Glymour, C., Lennox, J. G., Machamer, P., MsGuire, J. E., Norton, J. D., Salmon, W. C., & Schaffner, K. F. (1992). *Introduction to the philosophy of science*. Indianapolis, IN: Hackett Publishing Company.

Schlottmann, A., & Anderson, N. H. (1994). Children's judgments of expected value. *Developmental Psychology*, *30*(1), 56-66.

Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382-389.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, *43*(5), 1124-1139.

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., & Jenkins, A. C. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*(2), 211-223.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162-176.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322-332.

Scofield, J., & Behrend, D. A. (2008). Learning words from reliable and unreliable speakers. *Cognitive Development*, *23*(2), 278-290.

Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did she jump because she was the big sister or because the trampoline was safe? Causal inference and the development of social attribution. *Child Development*, *84*(2), 443-454.

Shanks, D. R. (1995). Is human learning rational? *The Quarterly Journal of Experimental Psychology*, *48*(2), 257-279.

Siler, S. A., & Klahr, D. (2012). Detecting, classifying, and remediating: Children's explicit and implicit misconceptions about experimental design. In R.W. Proctor, E.J. Capaldi (Eds.), *Psychology of science: Implicit and explicit processes* (pp. 137-180). New York, NY: Oxford University Press.

Silverman, I., Shulman, A. D., & Wiesenthal, D. L. (1970). Effects of deceiving and debriefing psychological subjects on performance in later experiments. *Journal of Personality and Social Psychology*, *14*(3), 203-212.

Skiba, R. J., Chung, C. G., Trachok, M., Baker, T. L., Sheya, A., & Hughes, R. L. (2014). Parsing disciplinary disproportionality: Contributions of infraction, student, and school characteristics to out-of-school suspension and expulsion. *American Educational Research Journal*, *51*(4), 640-670.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*(4), 753-766.

Spelke, E. (1994). Initial knowledge: Six suggestions. *Cognition*, *50*(1), 431-445.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605-632.

Spielberg, J. M., Galarce, E. M., Ladouceur, C. D., McMakin, D. L., Olino, T. M., Forbes, E. E., Silk, J. S., Ryan, N. D., & Dahl, R. E. (2015). Adolescent development of inhibition as a function of SES and gender: Converging evidence from behavior and fMRI. *Human Brain Mapping*, *36*(8), 3194-3203.

Stanovich, K. (2011). *Rationality and the reflective mind*. New York, NY: Oxford University Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453-489.

Talbott, W. (2008). Bayesian Epistemology. In *The Stanford encyclopaedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian

Taylor, M., Cartwright, B. S., & Bowden, T. (1991). Perspective taking and theory of mind: Do children predict interpretive diversity as a function of differences in observers' knowledge? *Child Development*, *62*(6), 1334-1351.

Taylor, K. M., & Shepperd, J. A. (1996). Probing suspicion among participants in deception research. *American Psychologist*, *51*(8), 886-887.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.

Todd, B.K., Barry, J.A., & Thommessen, S.A.O. (2017). Preference for 'gender-typed' toys in boys and girls aged 9 to 32 months. *Infant and Child Development*, 26(3), e1986.

Vanderbilt, K. E., Liu, D., & Heyman, G. D. (2011). The development of distrust. *Child Development*, *82*(5), 1372-1380.

Vanderborght, M., & Jaswal, V. K. (2009). Who knows best? Preschoolers sometimes prefer child informants over adult informants. *Infant and Child Development*, *18*(1), 61-71.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development*, *88*(1), 229-246.

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology*, *102*(1), 43-53.

Wente, A. O., Kimura, K., Walker, C. M., Banerjee, N., Fernández Flecha, M., MacDonald, B., Lucas, C., & Gopnik, A. (2017). Causal learning across culture and socioeconomic status. *Child Development*, *90*(3), 859-875.

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, *142*(4), 1006-1014.

Yamamoto, N., & Imai-Matsumura, K. (2019). Gender differences in executive function and behavioural self-regulation in 5 years old kindergarteners from East Japan. *Early Child Development and Care*, *189*(1), 56-67.

Yeo, J., & Tan, S. C. (2010). Constructive use of authoritative sources in science meaning-making. *International Journal of Science Education*, *32*(13), 1739-1754.

# Appendix A LETTERS FOR SCHOOLS

## SCHOOL RECRUITMENT EMAIL FOR SCHOOL A & B

To whom it may concern,

I would be really grateful if you could pass my request to your head teacher, Ms. XXX.
Thank you very much for your help,
Germaine Symons

Dear Ms XXX,
I am doing some research looking at how children reason about causal systems, and was wondering if I could work with the children at XXX?

I am CRB checked, and my study has the appropriate ethical approval. In fact, the children with whom I have already done this study all participated happily, and seemed to enjoy what they were doing.

I have attached an information sheet with more details regarding the study.

I would be very happy to visit you to discuss this further if you wish. Alternatively, my phone number is XXX XXXX.

Thank you for your time,

Germaine

Germaine Symons,
Department of Psychological Sciences,
Birkbeck University of London,
Malet St, London WC1E 7HX

TEL: 020 7079 0868
MOB: XXX XXXX
EMAIL: g.symons@bbk.ac.uk

INFORMATION FOR HEAD TEACHER ATTACHED TO RECRUITMENT EMAIL.

DEPARTMENT OF PSYCHOLOGICAL SCIENCES
BIRKBECK UNIVERSITY OF LONDON

Title of Study: Understanding causal systems in children

Name of researcher: Germaine Symons,
Department of Psychological Sciences,
Birkbeck University of London,
Malet St,
London WC1E 7HX
TEL: 020 7079 0868
MOB: XXX XXXX
EMAIL: g.symons@bbk.ac.uk

We are interested in how children learn about causal systems in everyday life, and through the teaching of science. In particular, we are interested in how children incorporate information about the causal system into their understanding, when the information comes from different sources (e.g. parents, teachers, other children, and so on); and how this changes as children get older and become more verbally competent. An understanding of the interaction between the utilization of different sources of information and age can contribute towards developing more effective methods of teaching of science in school.

We would like to test all Year 2, Year 4, and Year 6 children who consent, and whose parents have consented, to their participation.

The apparatus will be a game with slopes of different surface friction, angle, and starting position. A car is run down the slopes and, depending on the particular set up (type of surface, angle of slope, starting position on the slope), will travel a certain predictable distance. This apparatus represents the causal system we will use to assess children understanding.

We would ask children to 'play' with the game, and then ask them questions relating to how far they think a car will travel. The children would do this in three sessions. The first session would familiarise them with the apparatus, and establish what they already know about how the system works. The second introduces new information, and then looks at how this information might change their understanding. These first two sessions will be only a few days apart. The third session will occur a few weeks later and allows us to measure the effect of consolidation of knowledge on understanding.

Should you be interested, I will provide you with a summary of my findings that can be distributed once the study is complete, and/or do a presentation of the findings for anyone who wishes to attend.

The study is supervised by Professor Mike Oaksford. If you wish to contact my supervisor, contact details are: Department of Psychological Sciences, Birkbeck University of London, Malet St, London WC1E 7HX. TEL: 020 7079 0868

# Appendix B LETTERS FOR PARENTS

## INFORMATION SHEET FOR PRIMARY SCHOOL PARENTS

Dept. of Psychological Sciences
Birkbeck, University of London
Malet Street,
London, WC1E 7HX
www.bbk.ac.uk/psyc

| Researchers | Email | Telephone |
|---|---|---|
| Germaine Symons | g.symons@bbk.ac.uk | XXX XXXX or 020 7079 8008 |
| Prof. Mike Oaksford | m.oaksford@bbk.ac.uk | 020 7079 0868 |
| Prof. Andy Tolmie | a.tolmie@ioe.ac.uk | 020 7612 6888 |


Dear Parent / Caregiver,

I am doing a PhD based at Birkbeck, University of London, and work as a developmental psychologist. We are starting a project which looks at how children learn about causal systems in everyday life, and through the teaching of science. In particular, we want to know how information children have received from different people (teachers, parents, other children) affects how they understand how the world works. This will both help us understand children's reasoning about the world, and allow us to develop more varied and effective ways of teaching science in schools. XXX School, has very kindly agreed for us to be involved in this project, and I am writing to ask for your permission for your child to be included as a participant in this project.

Firstly, your child would play a language game, where they match spoken words to pictures. After that they get to play a game, where they run cars down different types of slopes. We then ask them questions about how far they think the car will travel. Your child would do this once where they get to play with the game under the guidance of the experimenter, and twice under their own guidance. Each session takes approximately 15 minutes. Children generally find the games fun and enjoy discussing the different things that might affect how far the car travels. We would collect our data by videoing your child's responses. These

video files will be kept in password-protected files, and only accessed by the researcher. All information collected will be kept confidential.

For your reassurance, all researchers involved in this investigation have been through the formal CRB Disclosure procedure and have been approved by the Birkbeck, University of London, to work with children. They also have previous experience working with children, both as a researcher and in a school. The project has been reviewed and approved by Birkbeck, University of London, Ethics and Research Committee. We would also like to emphasise that participation is entirely voluntary and children are free to withdraw from the study at any time. All parents will be provided with a summary of our findings (through the school) once the study is complete.

We very much hope that you and your child will offer to help us with our research. If you have any questions, please **contact Germaine Symons by phone or email** (see above). If your child and you are happy to take part, please sign the attached consent form and return it to your class teacher.

Many thanks,

Germaine Symons

CONSENT FORM FOR PRIMARY SCHOOL PARENTS

Dept. of Psychological Sciences
Birkbeck, University of London
Malet Street,
London, WC1E 7HX
www.bbk.ac.uk/psyc

| Researchers | Email | Telephone |
|---|---|---|
| Germaine Symons | g.symons@bbk.ac.uk | XXX XXXX or 020 7079 8008 |
| Prof. Mike Oaksford | m.oaksford@bbk.ac.uk | 020 7079 0868 |
| Prof. Andy Tolmie | a.tolmie@ioe.ac.uk | 020 7612 6888 |


Consent Form

- I consent to my son/daughter's involvement in this study
- I understand that my son/daughter's participation in this study is voluntary and that I may withdraw them at any point.
- I understand that all personal information will remain confidential to the Investigators.
- I have received a copy of the information sheet about this study.
- I have been given the opportunity to ask any questions that I may have about the study and have had these answered to my satisfaction.

Name of Child:_____

Date of birth of Child:_____

Parent/ Guardian's signature:_____

Researcher's signature:_____

Date: _____

Dept. of Psychological Sciences
Birkbeck, University of London
Malet Street,
London, WC1E 7HX
www.bbk.ac.uk/psyc

| Researchers | Email | Telephone |
|---|---|---|
| Germaine Symons | g.symons@bbk.ac.uk | XXX XXXX or 020 7079 8008 |
| Prof. Mike Oaksford | m.oaksford@bbk.ac.uk | 020 7079 0868 |
| Prof. Andy Tolmie | a.tolmie@ioe.ac.uk | 020 7612 6888 |

Dear Parent / Caregiver,

We are doing research based at Birkbeck, University of London, and work as developmental psychologists. We are doing a project which looks at how children learn about causal systems in everyday life, and through the teaching of science. In particular, we want to know how information children have received from different people (teachers, parents, other children) affects how they understand how the world works. This will both help us understand children's reasoning about the world, and allow us to develop more varied and effective ways of teaching science in schools. We have done this in primary schools, and now want to do it in secondary schools. XXX School has very kindly agreed for us to be involved in this project, and we are writing to ask for your permission for your child to be included as a participant.

Firstly, your child would do a language game, where they match spoken words to pictures. After that they get to play a game, and we ask them questions about how the game works. Your child would do this twice where they get to play with the game under their own guidance, and once under the guidance of the experimenter. Each session takes approximately 10-15 minutes, and would take place during lunch or after school.

Both children and adults generally find the games fun and enjoy discussing the different things that might affect how far the car travels. In return for participating, your child would be given a £5 gift voucher.

We would collect our data by videoing your child's responses. These video files will be kept in password-protected files, and only accessed by the researcher. All information collected will be kept confidential.

For your reassurance, all researchers involved in this investigation have been through the formal DBS Disclosure procedure and have been approved by the Birkbeck, University of London, to work with children. They also have previous experience working with children, both as a researcher and in a school. The project has been reviewed and approved by Birkbeck, University of London, Ethics and Research Committee. We would also like to emphasise that participation is entirely voluntary and children are free to withdraw from the study at any time. All parents will be provided with a summary of our findings (through the school) once the study is complete.

We very much hope that you and your child will offer to help us with our research. If you have any questions, please **contact Germaine Symons by phone or email** (see above). If your child and you are happy to take part, please sign the attached consent form and return it to Ms Budd (Psychology and Biology teacher).

Many thanks,

Germaine Symons

Dept. of Psychological Sciences
Birkbeck, University of London
Malet Street,
London, WC1E 7HX
www.bbk.ac.uk/psyc

| Researchers | Email | Telephone |
|---|---|---|
| Germaine Symons | g.symons@bbk.ac.uk | XXX XXXX or 020 7079 8008 |
| Prof. Mike Oaksford | m.oaksford@bbk.ac.uk | 020 7079 0868 |
| Prof. Andy Tolmie | a.tolmie@ioe.ac.uk | 020 7612 6888 |

Consent Form for Parents/Caregivers

- I consent to my son/daughter's involvement in this study
- I understand that my son/daughter's participation in this study is voluntary and that I may withdraw them at any point.
- I understand that all personal information will remain confidential to the Investigators.
- I have received a copy of the information sheet about this study.
- I have been given the opportunity to ask any questions that I may have about the study and have had these answered to my satisfaction.

Name of Child:_____

Date of birth of child: _____ Yr Group & Class _____

Parent/ Guardian's signature:_____

Researcher's signature:_____

Date: _____

# Appendix C INFORMATION FOR STUDENTS OVER 16

## CONSENT FORM FOR SECONDARY SCHOOL CHILDREN UNDER 16

Dept. of Psychological Sciences
Birkbeck, University of London
Malet Street,
London, WC1E 7HX
www.bbk.ac.uk/psyc

| Researchers | Email | Telephone |
|---|---|---|
| Germaine Symons | g.symons@bbk.ac.uk | XXX XXXX or 020 7079 8008 |
| Prof. Mike Oaksford | m.oaksford@bbk.ac.uk | 020 7079 0868 |
| Prof. Andy Tolmie | a.tolmie@ioe.ac.uk | 020 7612 6888 |


Consent Form (for students less than 16 years old – Yr 8 & 10)

- I consent to participating in this study
- I understand that my participation in this study is voluntary and that I may withdraw at any point.
- I understand that all personal information will remain confidential to the Investigators.
- I have received a copy of the information sheet about this study.
- I have been given the opportunity to ask any questions that I may have about the study and have had these answered to my satisfaction.

Name: _____

Date of birth: _____ Yr Group & Class _____

Signature: _____

Researcher's signature: _____

Date: _____

Dept. of Psychological Sciences
Birkbeck, University of London
Malet Street,
London, WC1E 7HX
www.bbk.ac.uk/psyc

| Researchers | Email | Telephone |
|---|---|---|
| Germaine Symons | g.symons@bbk.ac.uk | XXX XXXX or 020 7079 8008 |
| Prof. Mike Oaksford | m.oaksford@bbk.ac.uk | 020 7079 0868 |
| Prof. Andy Tolmie | a.tolmie@ioe.ac.uk | 020 7612 6888 |

Dear Student,

We are doing research based at Birkbeck, University of London, and work as developmental psychologists. We are doing a project which looks at how children learn about causal systems in everyday life, and through the teaching of science. In particular, we want to know how information children have received from different people (teachers, parents, other children) affects how they understand how the world works. This will both help us understand children's reasoning about the world, and allow us to develop more varied and effective ways of teaching science in schools. We have done this in primary schools, and now want to do it in secondary schools. XXX School has very kindly agreed for us to be involved in this project, and we are hoping you would like to be included as a participant.

Firstly, you would do a language game, where you match spoken words to pictures. After that you get to play a game, and we will ask you questions about how the game works. You would do this twice once when you would play with the game under your own guidance; and once under the guidance of the experimenter. Each session takes approximately 10-15 minutes, and would take place during lunch or after school.

Both children and adults generally find the games fun and enjoy discussing the different things that might affect how far the car travels. In return for participating, you would be given a £5 gift voucher.

We would collect our data by videoing your responses. These video files will be kept in password-protected files, and only accessed by the researcher. All information collected will be kept confidential.

For your reassurance, all researchers involved in this investigation have been through the formal DBS Disclosure procedure and have been approved by the Birkbeck, University of London, to work with children. They also have previous experience working with children, both as a researcher and in a school. The project has been reviewed and approved by Birkbeck, University of London, Ethics and Research Committee. We would also like to emphasise that participation is entirely voluntary and you are free to withdraw from the study at any time. You will be provided with a summary of our findings (through the school) once the study is complete.

We very much hope that you will offer to help us with our research. If you have any questions, please **contact Germaine Symons by phone or email** (see above). If you are happy to take part, please sign the attached consent form and return it to XXX

Many thanks,

Germaine Symons

Dept. of Psychological Sciences
Birkbeck, University of London
Malet Street,
London, WC1E 7HX
www.bbk.ac.uk/psyc

| Researchers | Email | Telephone |
|---|---|---|
| Germaine Symons | g.symons@bbk.ac.uk | XXX XXXX or 020 7079 8008 |
| Prof. Mike Oaksford | m.oaksford@bbk.ac.uk | 020 7079 0868 |
| Prof. Andy Tolmie | a.tolmie@ioe.ac.uk | 020 7612 6888 |

Consent Form (for students over 16 years old – Yr 12)

- I consent to participating in this study
- I understand that my participation in this study is voluntary and that I may withdraw at any point.
- I understand that all personal information will remain confidential to the Investigators.
- I have received a copy of the information sheet about this study.
- I have been given the opportunity to ask any questions that I may have about the study and have had these answered to my satisfaction.

Name: _____

Date of birth: _____ Yr Group & Class _____

Signature: _____

Researcher's signature: _____

Date: _____